Machine learning based Descriptive Statistical Analysis on Google Play Store Mobile Applications

Palagati Bhanu Prakash Reddy
Software developer
ValueLabs LLP
Hyderabad, India
bhanupalagati@gmail.com

Ramesh Nallabolu
System Engineer
Infosys Limited
Mysore, India
ramesh97.q1@gmail.com

Abstract-- **Handheld devices are vital for communication and went through a lot of advancements since the inception of this technology. Apart from normal GSM calling, communicating with each other is being done in a lot of ways with the help of the internet and the respective operating system application stores of the mobile phone. Developers and users are becoming ubiquitous for creating and using these applications, and Google play store is having a big share of 86.1% by 2019. So, performing a detailed analysis of the Google play store will help the developers to take reasonable decisions while picking the features and categories. To perform this analysis the "Google-Playstore-Full" dataset created by Gautham Prakash is used, and it has ~2,67,000 apps data that were crawled from the google play store, this dataset is available on Kaggle data world for public use. This analysis will provide insights about the correlation between various attributes of the applications moreover, it answers the question "Can Machine-Learning be performed on this dataset with the available attributes?" If can't then what are the best attributes to perform such an operation.**

Keywords -- Descriptive Statistics, Correlation, Categorical comparison of Mobile Applications, Google Play Store, Machine Learning

I. INTRODUCTION

As everyone knows, the hand-held devices market is colossal. It is no surprise that it attracted a great deal of competition. The following four are the top application stores for mobile phones, they are Google Play store with 2.57 Million apps, Apple store with 1.84 Million apps, Microsoft store with 0.669 Million apps, and Amazon Appstore with 0.489 Million apps at the end 2019 fourth quarter [1]. It is worth noting that the Google Play store itself has a greater number of apps than the Apple and Microsoft app stores. Numerous reasons are responsible for this result, the most important being the relative ease with which a developer can create an application and publish it in the Google Play Store. So, let's study the dynamics of Google Play store to get valuable insights that would help new developers while choosing various features or even the app domain for a new application.

Google Play store users and developers are not satiated with what they have on the store already. More and more apps are hitting the store, in the same way, more and more people are trying new apps. Just to mention the rate of change- on an average, 0.132 Million new apps [2] get added to the play store every month. If juxtapose this with the rate of change in the Amazon store, one can infer that Google play store can surpass the Amazon store in four months. That is both exhilarating and daunting at the same time.

As there is a non-trivial surge in the number of apps every day, can be seen as a promising future for an application and a developer. However, because of the sheer number of applications getting uploaded every day, the application reach may decline. Albeit, one can tackle this situation if he or she has some knowledge of the Google Play store app dynamics.

In this research article, a scrapped Google Play store dataset is going to be used available in the Kaggle data world. This is an eccentric data set because it has 0.27 Million apps data which is 10% of the humongous Google Play store [3]. So, can be safely presumed that the results of this sample are going to be the nearest simulation of the entire population.

In this dataset got 11 attributes for each application they are *app name, category, rating, reviews, installs, size, price, content rating, last updated, minimum android version, latest app version* [3]. All of them are very informative to analyze trends in the Google Play store. However, to make a descriptive analysis only needed category, ratings, reviews, installs, size, and price. Let's add another attribute called paid which holds 1 if an application is paid, else it holds 0.

Every dataset requires some pre-procession and the one are being worked with is no exception. The installs attribute has '+' symbol suffixing every value i.e., 1000000+ to depict a million-plus download. As have no way to mine the exact number let's get rid of this '+' and replace missing values with the mean of available values. Then size is suffixed with M or K to delineate whether the value is in MB or KB. So, get rid of M and convert values with K to M and get rid of K. Moreover, convert all the fields except categories and paid to numbers. As a result of the above pre-processing, an analysis-ready dataset is available[4].

II. LITERATURE REVIEW

Most of the analysis papers are only talking about the specific applications in the play store or the applications pertinent to an organization [5][6]. To be specific, they are dealing mostly with the sentiment analysis of an application based on the user-provided reviews [7][8], these kinds of researches are paramount to an organization and an application[9]. However, a holistic analysis of the platform will reveal the intricacies of the platform. This is beneficial to the developer, user, and even for the platform. For instance, if one category is seen no growth although other categories in the same position are showing exponential growth will alert the platform because the plausible reason is developers are having a hard time creating new applications in that category. Similarly, on seeing these analysis developers can find out the most appealing category of applications, as a result, they can create more of that type and attract more users. In the same way, users can see the unexplored categories and find out the most useful applications.

On the other hand, the existing research on the play store applications considers only 10,000 applications [10][11]. There are 50 categories in the Google play store [12]. Although the applications are evenly spread around the categories is considered, which is highly unlikely, each category will get only 200 applications and this data won't answer all the questions on the entire population because the data scraped on this level mostly incorporated popular applications so, this is highly biased. Along with that, if one category is high in number it will easily dominate the other categories.

III. METHODOLOGY

As the first step after collection and pre-processing it is very important to get clear insights into the data distribution, as a result, got mean, standard deviation, minimum and maximum value, quartiles 1, 2, and 3 of all the numeric attributes. Using this data, a lot of things can be presumed. This is a colossal dataset so; it is better to group all the applications based on the category and perform the analysis. To compare all categories with all the attributes under one roof the attributes will get standardized based on the min-max standardization. The reason for picking this over the Z score standardization is for min-max the resulting value is between 0 and 1. In contrast, for Z score had no control it can even range between -infinity to +infinity. After that, the data will be used to get the distribution of paid and free applications. Finally, the data will be used to find the correlation between all the attributes, this correlation matrix will give a clear idea i.e., a machine learning model can be created using these attributes or should mine for some other attributes. If unable to create a model then should mine for other valuable attributes. The flow chart named *Fig 1* depicts the process of analysis.

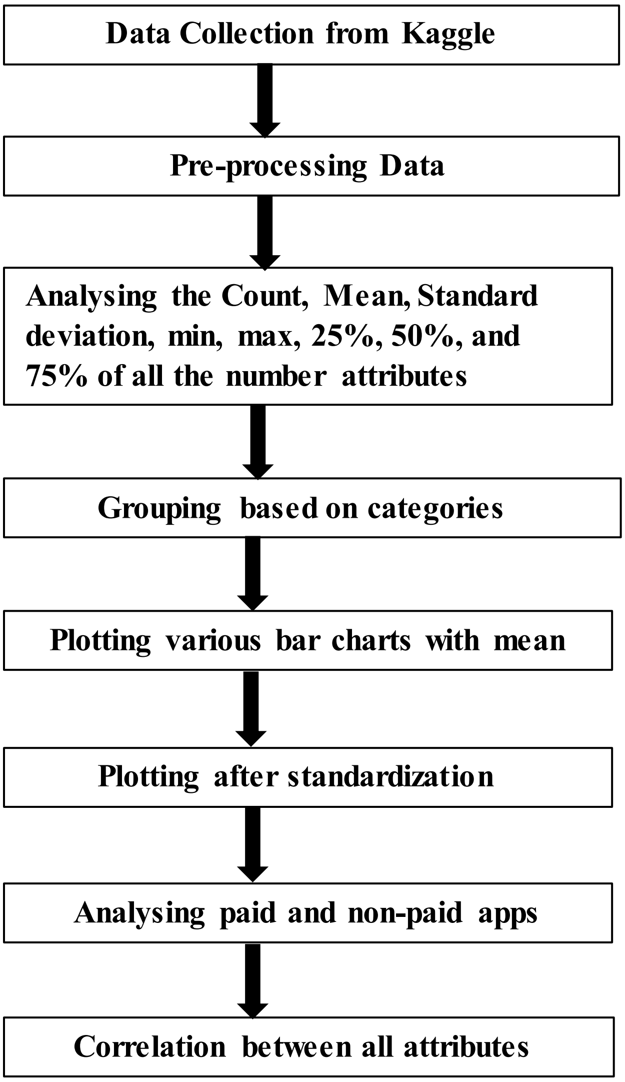

Data Collection from Kaggle
Pre-processing Data
Analysing the Count, Mean, Standard deviation, min, max, 25%, 50%, and 75% of all the number attributes
Grouping based on categories
Plotting various bar charts with mean
Plotting after standardization
Analysing paid and non-paid apps
Correlation between all attributes

Fig. 1. *Google play store analysis methodology diagram*

IV. RESULTS and DISCUSSION

A. *Statistical Description*

The statistical description is the first stage of the descriptive analysis. In this step, will be getting a holistic outlook of the data possessed. This analysis will give eight attributes for each column in the dataset. Those attributes are ***Count, Mean, Standard Deviation, Minimum, Maximum, and q1, q2, and q3 quartile ranges.***

$$\sigma(Std\ Deviation) = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \qquad \boldsymbol{eq.(1)}$$

Formula to calculate the standard deviation of the data

n = *number of values in a column* x_i = *ith value in the given column*

μ = *mean of all values in a column*   *eq.(2)*

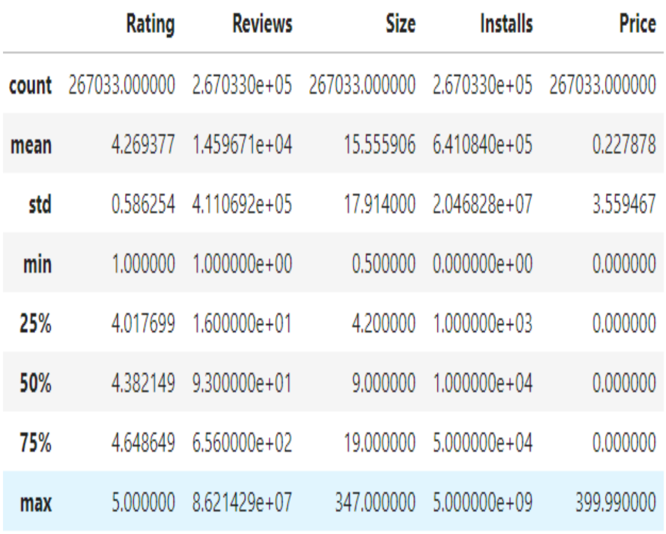
To get the proper intuition of the attributes and those values let's take the "**Ratings**" column and elucidate the values of the descriptive attributes.

$$mean = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \boldsymbol{eq.(2)}$$

Formula to calculate the mean of data

x_i = *ith value in the given column* n = *number of values in a column*

Firstly, had the ***count*** attribute this gives the number of values present in the "**Rating**". Secondly, got the ***Mean*** attribute that will provide the average value of all the values. For example, the "**Rating**" has a mean of 4.269377. Thirdly, had a ***standard deviation.*** *This* is a method of calculating variations in the data if the value is higher, then the spread of the data will also be higher. For "**Rating**" the ***standard deviation*** is 0.586254. Fourthly, there is a ***Minimum*** value and along with that, got ***Maximum*** value at the end. The ***minimum*** value of "**Rating**" is 1.0 and the maximum value of Ratings is 5.0. Finally, had ***q1, q2,*** and ***q3*** values of the column.



	Rating	Reviews	Size	Installs	Price
count	267033.000000	2.670330e+05	267033.000000	2.670330e+05	267033.000000
mean	4.269377	1.459671e+04	15.555906	6.410840e+05	0.227878
std	0.586254	4.110692e+05	17.914000	2.046828e+07	3.559467
min	1.000000	1.000000e+00	0.500000	0.000000e+00	0.000000
25%	4.017699	1.600000e+01	4.200000	1.000000e+03	0.000000
50%	4.382149	9.300000e+01	9.000000	1.000000e+04	0.000000
75%	4.648649	6.560000e+02	19.000000	5.000000e+04	0.000000
max	5.000000	8.621429e+07	347.000000	5.000000e+09	399.990000

Fig. 2. *Statistical descriptive analysis chart for the five attributes*

Source: A Result of the performed Analysis [13]

B. *Categorical Spread of Dataset*

Understanding the distribution of applications over the dataset is a vital step. This is to be performed at the beginning of the analysis.

The *Fig-3* named bar graph is a frequency distribution graph. At a glance through the graph, can be concluded that apps in the Education category are ubiquitous on play store ~ 33300 in the given dataset. On the other hand, Musical Games are very rare in the play store ~253 in the dataset.

There are a lot of independent educational institutions on the market and having an app and a website are the basic things they have to have. So, it is no surprise to see the colossal number of Educational applications. In contrast, there are very few musical instruments and most people use these applications as emulators to learn the real one. So, these applications are less in number.

However, all the statistical steps performed from now on are based on the grouping and averaging. So, even if the number of applications is less or humongous, it is not going to affect the analysis in any way.

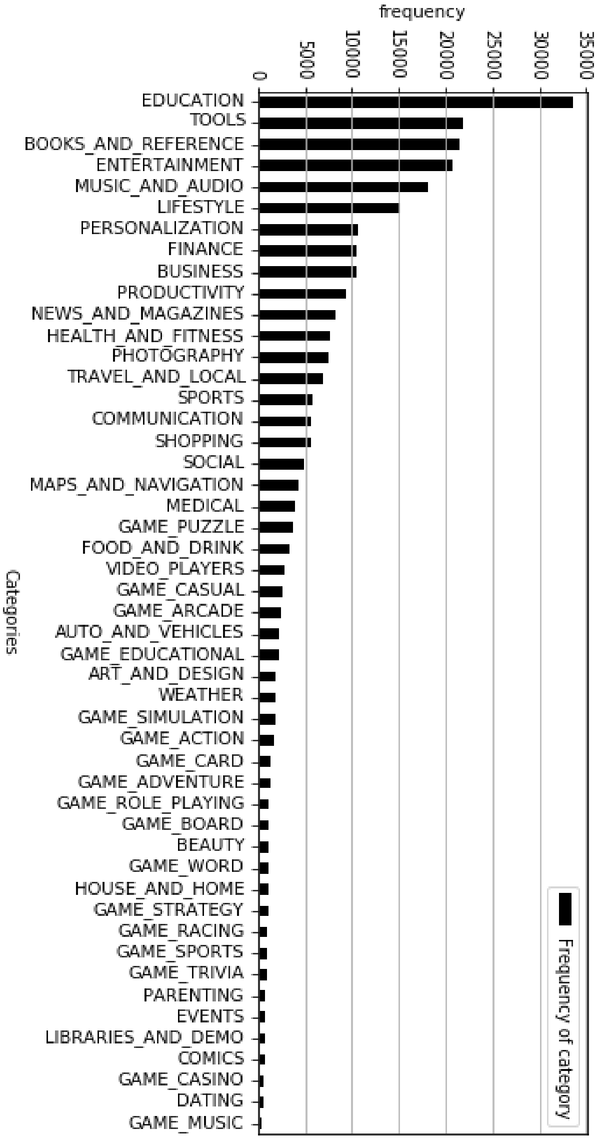


Fig. 3. *Bar chart depicting the categorical frequency of apps*
Source: A Result of the performed Analysis [13]

C. *Analysis of the categorical Ratings*

On the first glance of the *Fig-4* named bar chart, it is evident that all the categories have an average rating greater than 4.0. However, on further observing the graph can be seen that the **Books and Reference** category is having the highest rating i.e., ~4.5 and **Maps and Navigation** is the least performing with 4.0.

Let's see the plausible explanations for this user rating behavior. Although there are numerous reasons for this behavior, only a couple of these is eccentric. Firstly, not all countries have great internet connectivity and GPS coverage so, while travelling people often lose connectivity, as a result, they will lose their way. Secondly, most of the **Navigation** services are crowdsourcing. Sometimes there is a great possibility that the data were flawed. Because of the above two reasons, **Navigation** apps will often get low ratings. In contrast **Books and Reference,** applications are not labyrinth and user friendly.

D. *Analysis of the Categorical Installs*

The *Fig-5* named bar graph tells that the **Racing Game** category applications are surpassing every other category in installs with a significant difference. Contradictorily, the **Events** category is the least performing in terms of Installs.

Reasons for this behavior. As everyone knows, games are the most common installs. The graph strengthens that fact by listing 9 game categories in the top 11 because from kids to adults everyone was interested to play one or the other. To be specific, **Racing Games** were quick to set up, easy to play, and appealing to all levels of people with fewer restrictions, so on average a **Racing Game** is acquiring about 55,40,268 installs. In the **Events** category, applications are very specific to a sector.

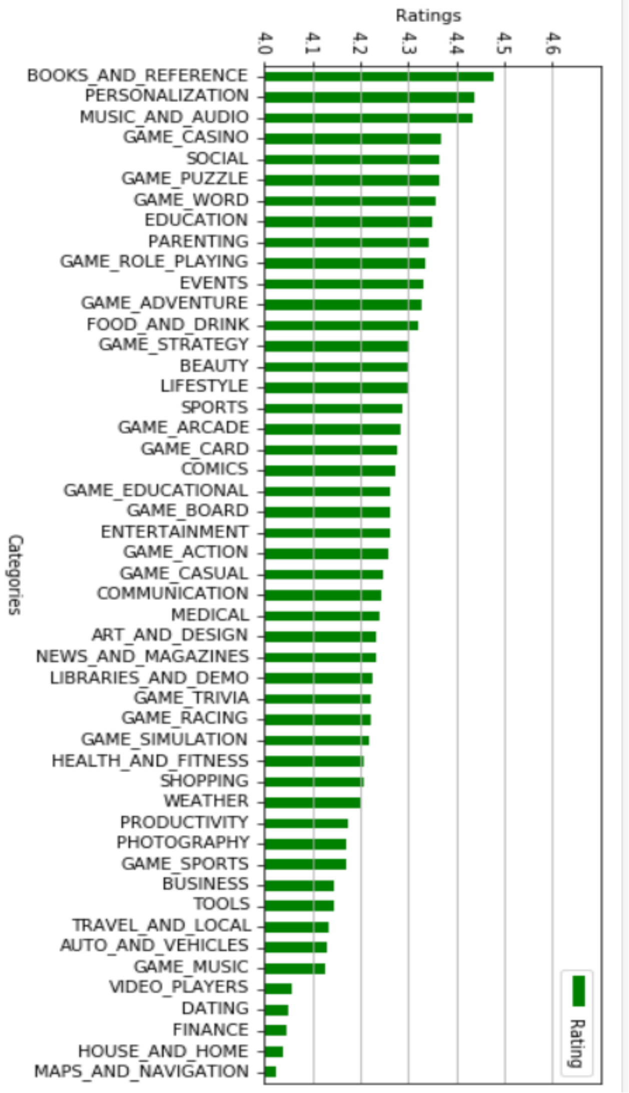


Fig. 4. *Bar chart depicting the categorical mean Ratings of apps*
Source: A Result of the performed Analysis [13]

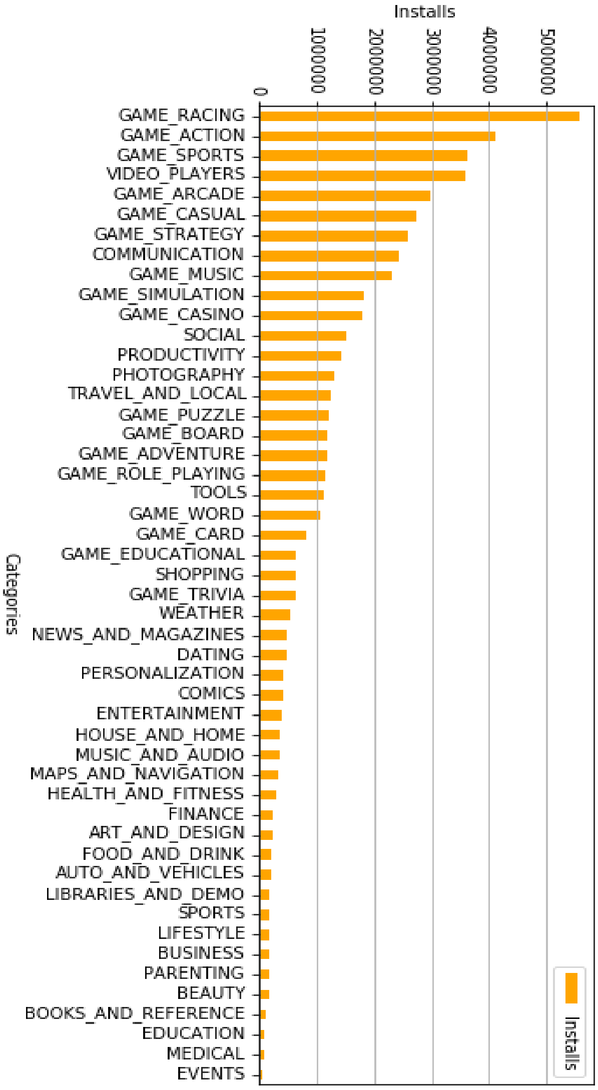


Fig. 5. *Bar chart depicting the categorical mean Installs of apps*
Source: A Result of the performed Analysis [13]

For example, if there is an application called the holiday calendar, it is plausible that this will be following the holiday schedule of just one country. As a result, the target users have greatly plummeted in this category. On average each **Event** category application is getting only 46,026 installs. In summary, the **Racing Games** category has got ~120 times more installs than an Events category application.

E. Analysis of the Categorical Reviews

The *Fig-6* named bar graph put forth that the **Action Games** are a highly reviewed category and **Strategy Games** were following it. On the other hand, **Events** are the least reviewed category in the play store.

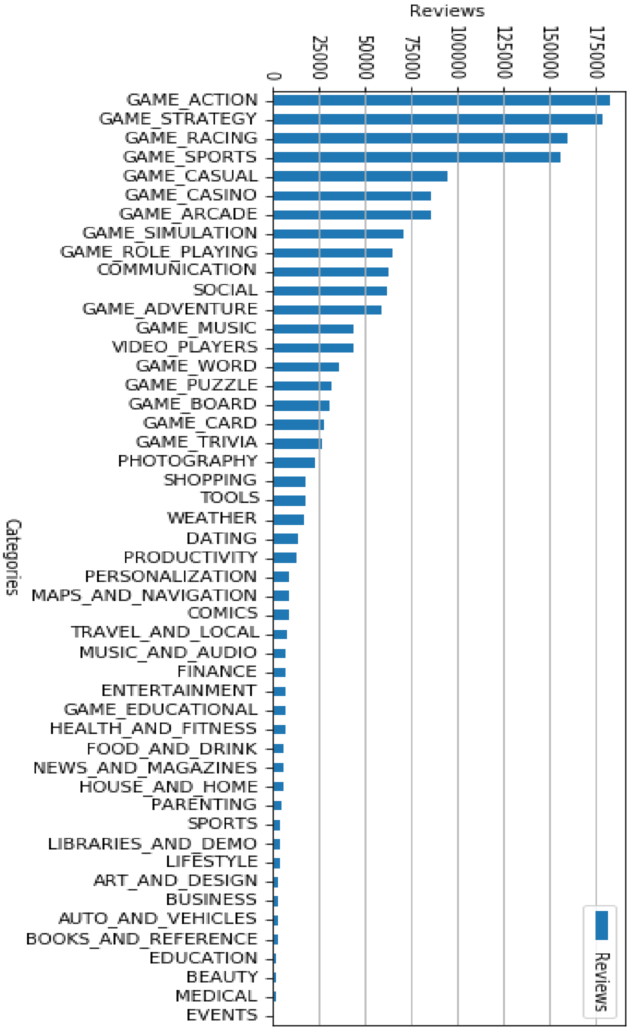


Fig. 6. *Bar chart depicting the categorical mean Reviews count of apps*
Source: A Result of the performed Analysis [13]

On analyzing the facts, the **Action** and **Strategy** games were mostly free-roaming and interactive games as well as it requires a lot of graphics, as a result, most users feel that there are lags and bugs and they use reviews to report their problems or appreciations i.e., 1,82,065 on average. In contrast, the number of installs for the **Events** category itself is very less, so it is no surprise to know that the number of reviews was also very few. To be specific, on average Events category application got 456 reviews which are very less.

F. Analysis of the Categorical Price

Based on the *Fig-7* named graph can be concluded that the **Role Play** and **Adventure Games** are highly-priced on average i.e., ~$1.3. On the other hand, **Events** are least priced with an average of ~$0.004.

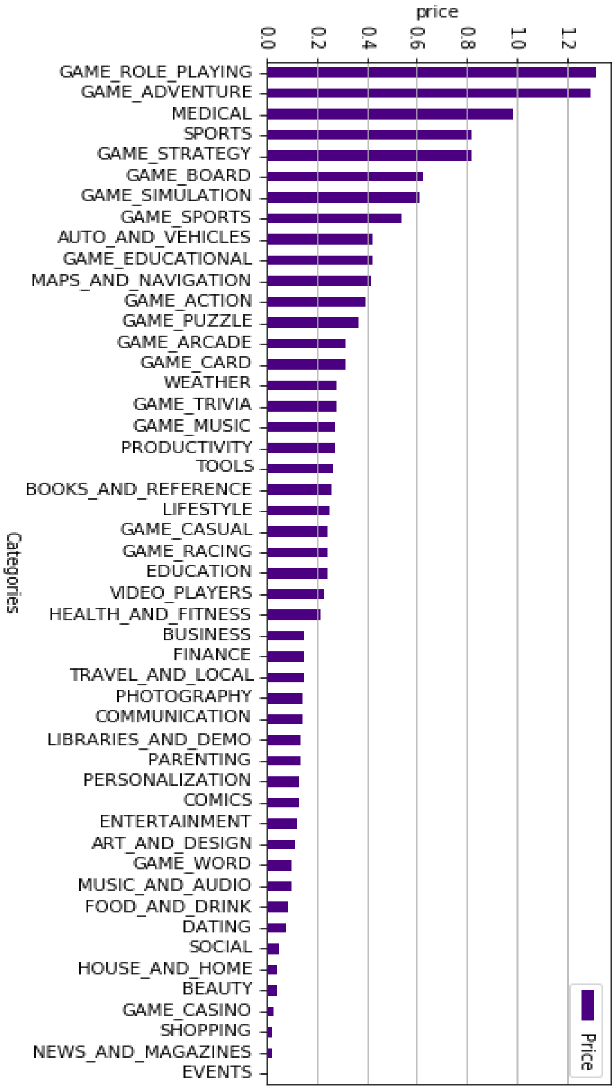


Fig. 7. *Bar chart depicting the categorical mean price of apps*
Source: A Result of the performed Analysis [13]

The above relation can be explained with the following reasons. Firstly, **Role Play** and **Adventure** games are free-roaming and highly graphic oriented games. So, it requires a lot of developers, designers, and testers to create games like the above. Besides, these games require a lot of bug fixes and maintenance because of the online play modes. In contrast, **Events** applications are fairly simple and there is no need for great teams unless there are complicated app synchronizations. Moreover, **Events** can be easily updated with a sophisticated backend mechanism which is a must for an app like that.

G. Analysis of the Categorical Size

On observing the *Fig-8* named bar chart, it is clear that games are the top size consuming applications. For instance, all top 15 categories in size are games and **Role Play** games were paramount with an average of ~48M.B. However, the least size consuming applications are Tools with an average of ~8M.B.

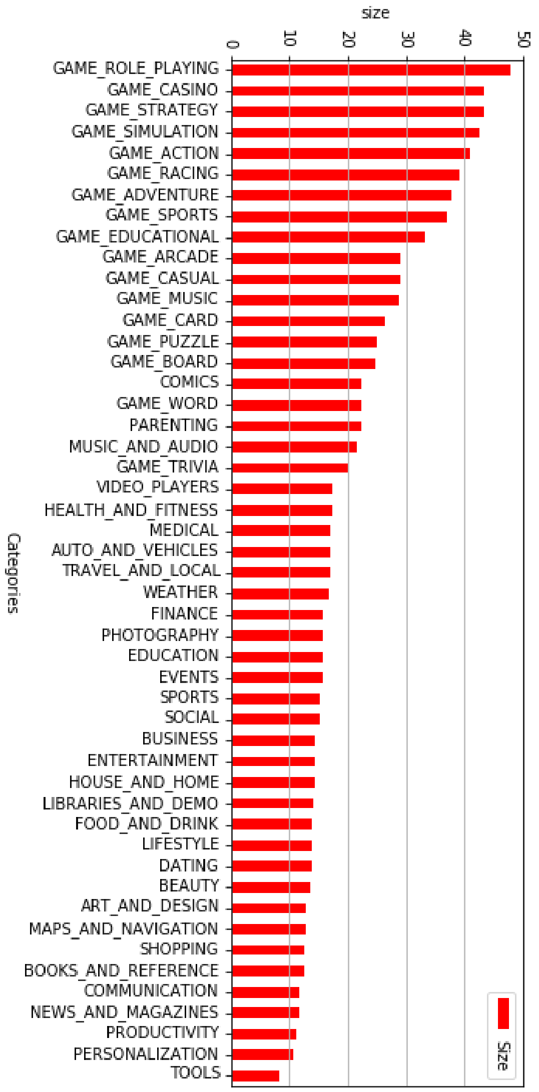


Fig. 8. *Bar chart depicting the categorical mean converted of apps*

Source: A Result of the performed Analysis [13]

A plausible explanation for this behavior is as follows. Games are filled with scenes and a scene is nothing but a sequence of properly arranged interactive 3D images. Since the assets are numerous the bundle size will be lofty. On the other hand, Tools category applications will use services provided by the operating system, so the lines of code are very few, on top of that, there is no huge requirement of assets. So, the size of the Tools category applications can be expected to be small.

H. Comparing the Paid and Free Apps

On seeing the *Fig-9* named pie chart, it is evident that most of the applications in the Google play store are free. To be precise, ~95.7% apps are free on play store and only ~4.3% apps are paid.

This brings us to two conclusions. Firstly, most of the application developers are following the "Freemium Business Model" in which most of their earnings are from promotions and advertisements[14]. Along with that, most of the users were also comfortable with the advertisements rather than putting a few coins into the developer pockets.

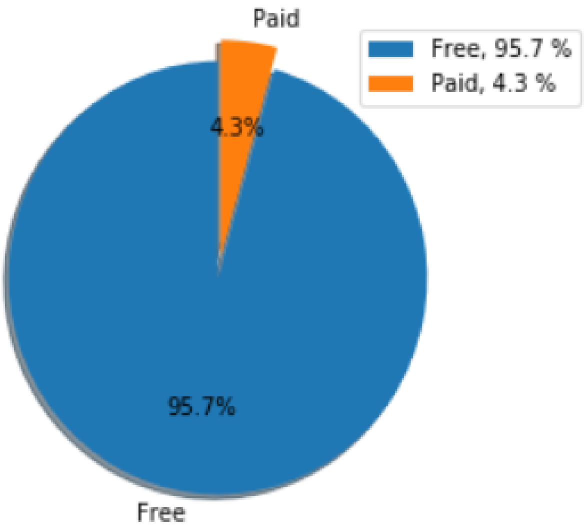


Fig. 9. *A pie chart depicting paid and non-paid apps spread on play store*

Source: A Result of the performed Analysis [13]

I. Analysis of the Scaled Categorical Values

This is an intuitive way of comparing multiple attributes of the data. The reason for this analysis is getting the performance of a category on a holistic picture.

These results are acquired by adding reviews, ratings, and installs and then subtracting the size and price. The addition and subtraction of the raw values of the attributes will result in an influenced value. For example, reviews and installs can go up to millions, but the ratings can only go up to 5.

Price and size are negative attributes so are subtracting them. The reason being if either the price or size is high for an application then the users are less likely to install that application if there is a counter application with the same quality and performance.

The possible lower limit of the scaled values is -2 when all the positive values are 0 and the negative values are 1. Similarly, the possible upper limit 3 when all the positive values are 1, and the negative values are 0.

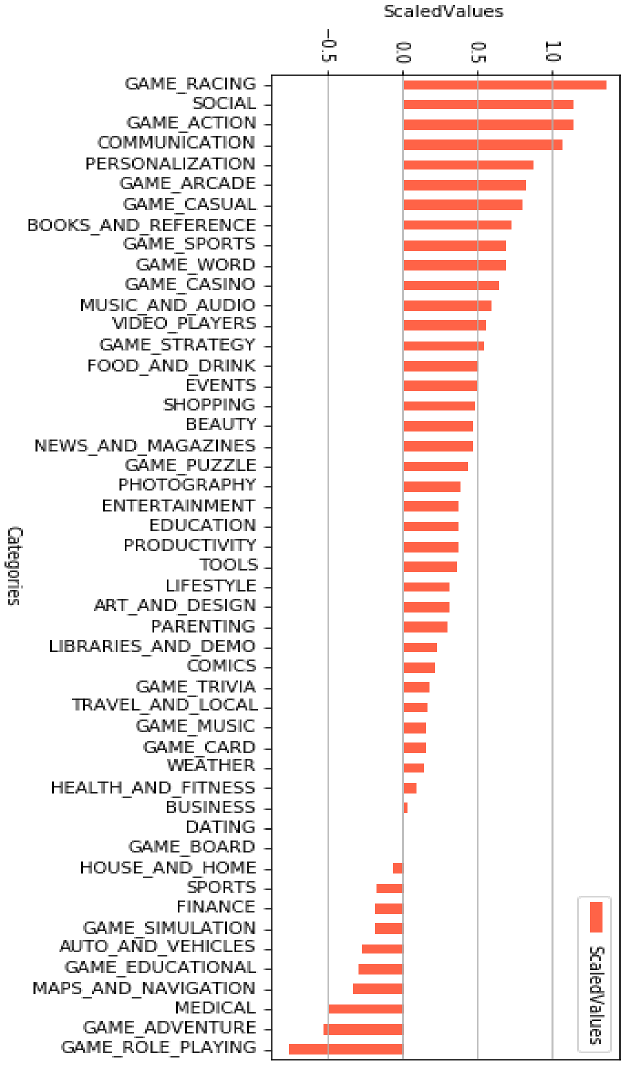


Fig. 10. *Bar chart depicting the scaled categorical values of apps*
Source: A Result of the performed Analysis [13]

From the *Fig-10* named graph, it is evident that the **Racing Games** are more likely to be tried by the users because of their low size and price. On the other hand, **Role Play** games are less likely to be tried by a plethora of users because of their high negative attribute values considering there is no external effect like marketing, existing movie or series, and so on.

$$MinMax = \frac{X - Minx}{Maxx - Minx} \qquad \boldsymbol{eq.}(\mathbf{3})$$

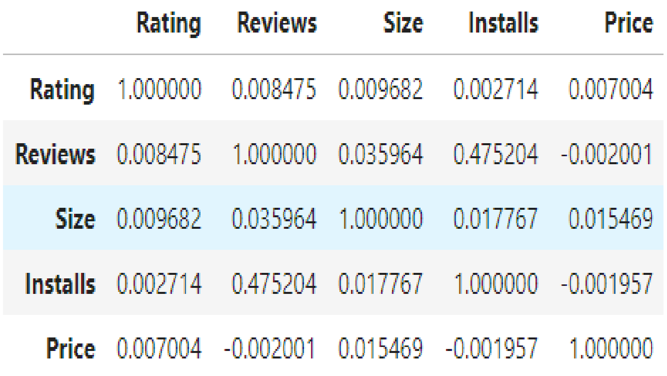
Min-max normalization formula

Minx = minimum value in the column
Maxx = maximum value in the column

X = current value
MinMax = Resultant value

$$Agg_i = Installs_i + Ratings_i + Reviews_i - Price_i - Size_i$$

J. Matrix of Coefficient of Correlation



	Rating	Reviews	Size	Installs	Price
Rating	1.000000	0.008475	0.009682	0.002714	0.007004
Reviews	0.008475	1.000000	0.035964	0.475204	-0.002001
Size	0.009682	0.035964	1.000000	0.017767	0.015469
Installs	0.002714	0.475204	0.017767	1.000000	-0.001957
Price	0.007004	-0.002001	0.015469	-0.001957	1.000000

Fig. 11. *Bar chart depicting the categorical frequency of apps*

Source: A Result of the performed Analysis [13]

The coefficient of correlation is an impeccable method to find out the relation between various columns in the given dataset. The values range from -1 to +1. If it is in the vicinity of -1 there is a negative correlation between the two attributes. If it is near to 0 there is no correlation. If it is near to 1 then there is a positive correlation.

If it is a negative correlation, an increase in one value decreases the other value. If it is a positive correlation, on the other hand, an increase in one value increases the other value.

$$r_{xy} = \sum_{i=1}^{n}((xi - \bar{x}) * (yi - \bar{y})) / \sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2} * \sqrt{\sum_{i=1}^{n}(xi - \bar{x})^2} \quad \boldsymbol{eq.}(\mathbf{4})$$

Coefficient of correlation formula

xi = ith value in the column X *yi = ith value in the column Y*
$\bar{x}$ *= mean of values in column X eq.(2)* $\bar{y}$ *= mean of values in column Y eq.(2)*

The above 5 X 5 matrix depicts all the correlation values. However, the upper triangle will only be looked at because the diagonal will depict the relation of one attribute with the same attribute and the lower triangle is the redundancy of the upper triangle.

On observing the above matrix, it is clear that Ratings does not correlate with any other attribute, Reviews attribute has a slight positive correlation with Installs ~ 0.475, Size and Price have no conspicuous relation with any other attribute in the dataset.

V. CONCLUSION

Based on the given dataset and the performed analysis it is evident that the *Ratings, Reviews, Price, Size, and Installs* have no strong correlation. Plausibly there will be other attributes need to gather to complete the analysis. Based on descriptive analysis education-related apps are ubiquitous on the play store and the musical games are rare. **Books and References** are highly rated and **Maps and Navigations** are least rated. **Racing Games** are frequently installed and **Events** are rarely installed. **Action games** are often reviewed and **Events** are rarely reviewed. **Roleplay games** are highly-priced and **Events** are less priced. **Role Play games** are humongous in size and **Tools** are small in size. On a holistic view, **Racing games** are best to take and **Role play games** are best to avoid.

Based on the results a developer can select a specific category based on his target requirements. This paper can be helpful for a developer to take valuable decisions that can change the reachability of an application.

VI. FUTURE WORK

Since the above dataset is not enough to create a machine learning model the future work should be focused on gathering the data that suits to create a machine learning model and to create one.

Plausibly the following data works much better rather than the available one.

1. Gathering how many times a user is using an application in a stipulated amount of time.
2. The rate of growth in application installations tells us the popularity and relevance of the application to vogue.
3. How much proportion of application features require in-app purchases?
4. How frequently a user is going to come across an advertisement.
5. Who is developing the application? If they have a group of users already using other applications, they can promote a new one easily rather than a new company or a developer.

If the application is coming from a movie, a physical game or something people already acquainted with will have an edge over the other applications.

On that note, any machine learning algorithm building on this dataset will result in an improper result because of ground touching correlation values.

REFERENCES

[1] Statistics of the top four Application stores in the mobile world https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/.

[2] Google Play Store daily and monthly new app upload statistics https://appinventiv.com/blog/google-play-store-statistics/.

[3] 267k scrapped Google play store dataset from the Kaggle data world.https://www.kaggle.com/gauthamp10/google-playstore-apps?select=Google-Playstore-Full.csv.

[4] Shi, Yong, and Daniel Brown. "An Attempt to Discover Analytical Information for Multi-Dimensional Data Sets." In 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1-5. IEEE, 2018.

[5] Fatema Akbar and Lusi Fernandez-Luque "What's in the store? A Review of Arabic Medical and Health Apps in the App Store" in 2016 IEEE International Conference on Healthcare Informatics.

[6] Dema Mathias Lumban Tobing, Ema Utami, and Hanif Al Fatta "Analysis of Dominants Game Elements Using the Sillaots Parameters and Octalysis Framework on the Google Play Store" in 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering.

[7] Soo Ling Lim and Peter J. Bentley "Investigating App Store Ranking Algorithm using a Simulation of Mobile App Ecosystems" in 2013 IEEE Congress on Evolutionary Computation June 20-23, Cancun, Mexico.

[8] Mir Riyanul Islam "Numeric Rating of Apps on Google Play Store by Sentiment Analysis on User Reviews" in International Conference on Electrical Engineering and Information & Communication Technology 2014.

[9] Golam Md. Muradul Bashir, Md. Showrov Hossen, Dip Karmoker, and Md. Junaeed Kamal "Android Apps Success Prediction Before Uploading on Google Play Store" in 2019 International Conference on Sustainable Technologies for Industry 4.0.

[10] Rana M. Amir Latif, M. Talha Abdullah, Syed Umair Aslam Shah, Muhammad Farhan, Farah Ijaz, and Abdul Karim in "Data Scraping from Google Play Store and Visualization of its Content for Analytics" in 2019 International Conference on Computing, Mathematics and Engineering Technologies.

[11] The Google play store analysis using 10,000 scrapped applications https://levelup.gitconnected.com/google-play-store-analysis-8501228a1ace.

[12] The available categories on the Google Play Store https://42matters.com/docs/app-market-data/android/apps/google-play-categories.

[13] Gist link of the performed descriptive analysis https://gist.github.com/bhanupalagati/d8760737290ee815729a343170e6f114.

[14] Zhiyong Li, Guofang Nan and Minqing Li in "Advertising or Freemium: The Impacts of Social Effects and service Quality on

Competing Platforms" in IEEE: Transactions on Engineering Management.