ANALYZING GOOGLE PLAY STORE APPS IN THE ANDROID MARKET

Sivasanggeri Balan (P-COM0161/18), **Khaw Siang Lee** (P-COM0023/19),
Thangeswariy Ramanei (P-COM0017/19), **Iswarya** (P-COM0165/18)
School of Computer Sciences,
11800, Universiti Sains Malaysia, Pulau Pinang, Malaysia.

Abstract - Google plays store apps get flooded with new apps daily regardless of sole developers or with teams or organizations or Google itself. This group of people working hard to make the apps successful with gigantic racing from all over the globe, it is important for a developer to know whether he is proceeding in the right direction. To understand the relationship between the Google Play Store Apps dataset, all the attributes are used to visualize them in Tableau. Thus, an app's success is usually determined by the number of installs and the user ratings that it has received over its lifetime rather than the profits it generated. In this project, we have tried to perform exploratory data analysis to dive deeper into the Google Play Store data that we obtained, discovering relationships with specific features such as paid, free, ratings, user review, category. and how features affect the installs, to use them to find out which apps are more likely to succeed. This visualization is more helpful for apps developers in the development phases, also for the users of the apps' application for the selection of the apps that they want to use.

1. Introduction

In the world of Big Data, data visualization tools and technologies are important to evaluate huge amounts of information and make data-driven decisions[1]. It is often that visualization displays are very efficient at communicating information[2]. When it comes to exploring data and presenting results to be viewed, the visualization shows a great approach. Visual elements such as charts, graphs, and maps provide an accessible way for data visualization tools to understand in detail the trends, outliers, and patterns of the data[1]. So, pretty much data visualization is a form of visual art that captivates our interest thus keeping us intrigued by the message it tryna convey[1]. The project involves explorations of how to carry out specific encoding and interaction ideas based on the application domain of the dataset that has been decided on. Tableau will be used to generate different types of visualization.

2. DOMAIN BACKGROUND

Google Play Store is the official marketplace for devices powered by the Android operating systems including smartphones, tablets, smart TVs, and numerous devices. 2.5 billion active Android devices drive the growth of the Android ecosystem. The number of applications published in the Google Play Store approaching 2.47 million as of the third quarter of 2019. It was first introduced in October 2008, formerly known as Android Market [3]. It shows that the Google play store become so widespread among the public and more developers are excited to design more apps. We selected the Google Play Store Apps data to find the success of the Google mobile app based on the overall rating of an app. Feedback and review are key elements to drive improvement for the apps. [4] In the introduction from Galvis and Winbladh, they do mention that user feedback is crucial to generate improvement in software quality. Most of the users trust the app by viewing the rating and reviews on it. More leading-rated apps are more likely to be recommended and presumed by users.

2.1 Domain Level

This visualization is mainly targeted at the users of the Google Apps Developer and business development unit. This visualization is aimed to identify the currently available application that is focused on which category, age population, and number of downloads. This is to provide an insight into the business development unit to shift focus to a more popular and profitable category and also provide insights to a developer to enhance their application based on user ratings and reviews.

3. DATA AND TASK ABSTRACTION

Author Munzer stresses abstraction level in his book which explains what dataset we dealt with and why the dataset is necessary. This covered both data and task abstraction. Many types of tools can be used to do this visualization. This level also incorporates the effect of using the particular tool in comparison with other tools that assist visualization of the data, and also addressing the tool and user's goal such as data representation and map the particular domain task [6].

3.1 Data Abstraction

Data abstraction is what type of data is to be used. So, for this project, we obtained data from Kaggle.com at the URL https://www.kaggle.com/lava18/google-play-store-apps, titled "Google Play Store Apps" which was originally extracted from Google Play Store by Lavanya Gupta. This data set consists of 10841 rows and 13 columns. Rating is the target that this project is predicting, and the rest are the features used for the prediction. The attribute type is interpreted as descriptive, numeric, nominal, and ordinal categorical and time series based on the description provided from the dataset and exploration of the first 30 rows from the dataset. The summary is tabled in Table 1.

No	Attributes	Description	Attribute Type	Feature/Class
1	App	Application name	Descriptive	Feature
2	Category	Category the app belongs to	Nominal Categorical	Feature
3	Rating	Overall user rating of the app	Numeric	Class
4	Reviews	Number of user reviews for the app	Numeric	Feature
5	Size	Size of the app	Numeric	Feature
6	Installs	Number of user installs for the app	Numeric	Feature
7	Type	Paid or Free	Boolean	Feature
8	Price	Price of the app	Numeric	Feature
9	Content Rating	Age group the app is targeted at - Children / Mature 21+ / Adult	Ordinal Categorical	Feature
10	Genres	An app can belong to multiple genres.	Nominal Categorical	Feature
11	Last Updated	Date when the app was last updated on Play Store	Time Series	Feature
12	Current Ver	Current version of the app available on Play Store	Ordinal Categorical	Feature
13	Android Ver	Min required Android version	Ordinal Categorical	Feature

3.2 Task Abstraction

This dataset provides insights into the Google Play Store Apps market from 2010 until 2018. Information visualization graphical illustration is required to understand the preferences apps by

category based on the reviews, the number of installs, price, and content rating. From the data given, data cleaning should be done since they are some missing values or unknown values. Not having a clean set of data could be reflected in the visualization. After the data preparation process, with the help of the visualization idioms, insights such as which App is mostly installed, which apps got more ratings, are the specific category is focused on can be seen clearly from the dataset.

3.2.1 Our task

i. Find the most downloaded and installed Category of Apps.
ii. Identify which Category has the most Apps.
iii. Identify the Distribution of user ratings.
iv. Identify the number of reviews in each Category.
v. Find the category which has the free/Paid Apps type.
vi. Identify the correlation between Installs and Reviews.
vii. Identify the correlation between rating and price.
viii List of all the Apps in Category with the rating.

4. RELATED WORK

Google plays store apps are one of the most popular apps used to download by the user and millions of apps uploaded by developers around the world. This domain has been analyzed by considering different datasets and parameters using different data analytics tools and techniques by many researchers.

One of the authors [5], used CIRCOS tool to visualize the dataset, especially using attributes Paid and Free application. CIRCOS is the tool in which multiple dependent and independent variables could be easily identified by the systematic point of view and best corporation of the layout. Ratings, In Application Purchases, Advertisements support, and Installs are being used for the analysis and better visualization of the attributes data of free and paid games applications by these authors. The author generated CIRCOS Visualization for Free Games with Four Attributes as below. The circular layout of the data could be easily identified and the position of the objects. The circular region has more advantages due to attractive responses. CIRCOS is the ideal creation for the publication quality and it also illustrates the high data into ink ratio. Which richly identifies the symmetries among all the values. The audience details and focuses on the point of the figured requirement of the image. For the visualization of the genomic data and the genomics, migration could be identified by the system data. The multi-layered mathematical art and the data describe the relationships of the multi-layered and annotations for the scales of CIRCOS as shown in the figure.

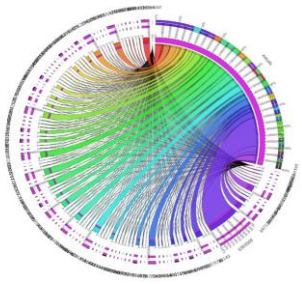

Figure: CIRCOS Visualization for Free Games with Four Attributes

So, it depends on what we needed from the visualization and what tools we used to generate what we needed. It fully depends on the data and task abstraction for the research purpose.

5. METHODS

Google play store app dataset with 8861 rows has been used in this study, to perform visualization analysis design.

Table 1: Data analysis idioms, tasks, justification, and figure number

No	Idioms	Task	Justification	Figure
1	Horizontal Bar Charts	Categorize Category of Apps and Type of Apps.	Filter some of the attributes, based on the overall category. Summarize the category distribution by Type of Apps	Fig. 1
2	TreeMap	Number of apps in each Category	Summaries total apps in each category.	Fig. 2
3	Histogram	Distributio n of user rating	Filter some attributes based on rating.	Fig.3
4	TreeMap	Number of reviews in each Category	Summaries total number of reviews in each Category.	Fig. 4
5	Stack Bar chart	Categorize the total of the app in each category according to paid and free App.	Summarize the total number of the app in each category based on the paid and free apps	Fig. 5

6	Scatter Plot	Categorize the relationship between Installs and Review	Filter Apps based on Reviews and Installs attribute. Show relationships between installs and reviews.	Fig. 6
7	Scatter Plots	To show the correlation between rating and price attributes	Summarize the correlation between attributes such as rating and prices.	Fig. 7
8	Gantt Chart	To list Apps in each Category with a rating	List all the Apps in each Category with a horizontal separation of Rating.	Fig.8

6. ANALYSIS

6.1 Strengths and Weakness for Idiom/Task

Table 2: Strength and weakness for idiom/task

Idiom/ Task	Strength	Weakness
1	**The vis shows the type of category and the distribution of the number of downloads in descending order. The game category (19.55%) has the highest number of downloads followed by tools and family while parenting, comics, and libraries have the lowest number of downloads.**	**Different Categories of Apps can be easily identified by free and paid apps.**
2	**User rating is heavily distributed between the rating 4.0 to 4.5. It is a negatively skewed distribution.**	**The category with the highest and lowest rating is not visible.**
3	**Different Categories of Apps can be easily identified by free and paid apps. The majority of the category has more free apps than paid apps. The family category has the highest number of free apps.**	**Different types of Apps cannot be easily identified by free and paid apps.**
4	**The treemaps show all the different types of available categories clearly. The heavily saturated categories have a higher number of reviews**	**The Apps related to each category cannot be identified.**
5	**Makes the outliers stand out**	**Each segment is difficult to be compared since they are not aligned on a common baseline**
6	**Since the dots are concentrated near the line, the relationship is considered to be strong**	**Does not show a relationship for more than two variables**
7	**Can manage to determine the outliers**	**Does not show a relationship for more than two variables**
8	**Easy to understand, clear, and visual representation of Apps based on Rating. It is easy to add dependencies and predecessors such as apps and rating dependencies.**	**It can be quite compact and hard to read and understand.**

6.2 Figures

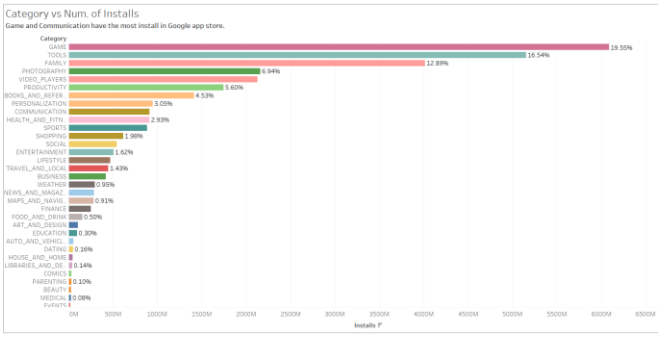

Figure 1. Horizontal Bar Charts showing Category of Apps and Type of Apps.

There are 33 categories of applications. The most popular apps are games, tools, family, and photography. The least popular apps are parenting, comics, dating, and education.

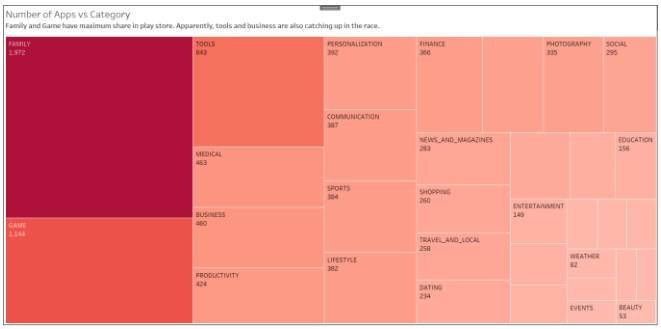

Figure 2. Treemap shows the number of apps in each category.

Family and Game categories have the highest number of App which are 1,972 and 1,144 respectively and followed by tools with 843. On the other hand, medical, business productivity, personalization, communication, sports, finance, photography, and lifestyle can be clustered into a range of 300 to 400 apps. Shopping, travel, entertainment fall below the range of 300 apps.

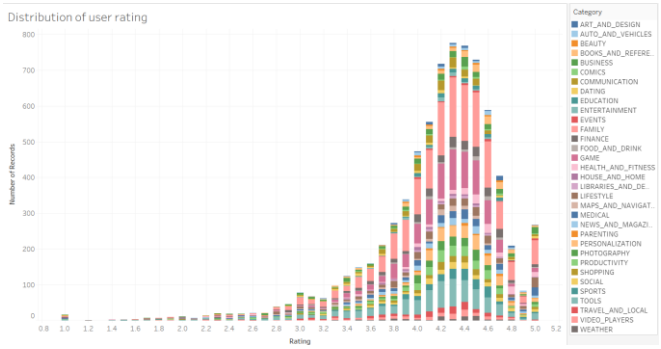

Figure 3. Stacked Bar Chart Histogram showing Distribution of User Rating

The user rating is positively skewed as the rating of the apps is mainly focused in the range 4.0 to 4.6. Family, games, and tools are highly rated in the range of 4.2 to 4.4. We also notice that the family has the highest rating of 5.0 with 67 records as well.

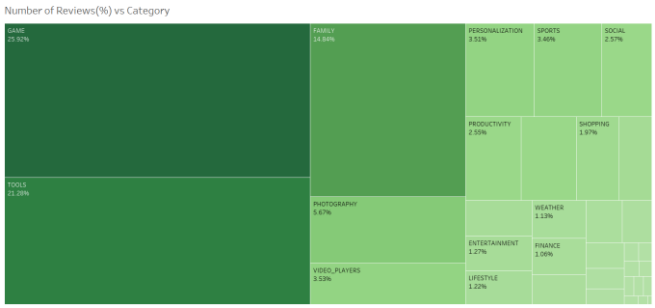

Figure 4. TreeMap shows the number of reviews in each Category.

The highest user reviews are cast to game, tools, and family with 25.92%, 21.28%, and 14.84% respectively. Lifestyle, finance, entertainment, and shopping have relatively low reviews by users.

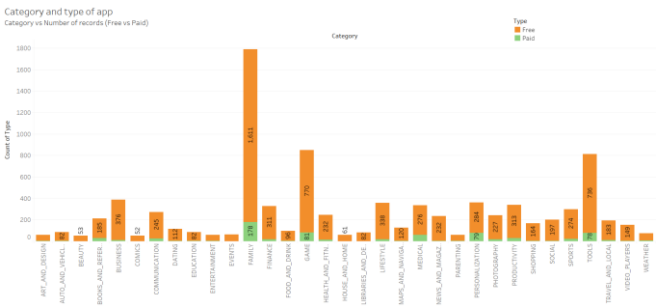

Figure 5. Stack Bar chart categorizes the total of the app in each category according to paid and free App.

The majority of the application across all categories are free of charge. The highest number of free applications are coming from family 1,611, games 770, and tools 736. The paid application is extremely minimal and is occurring only in books, communication, finance, game, lifestyle, medical, personalization, photography, productivity, sports, tools, and travel.

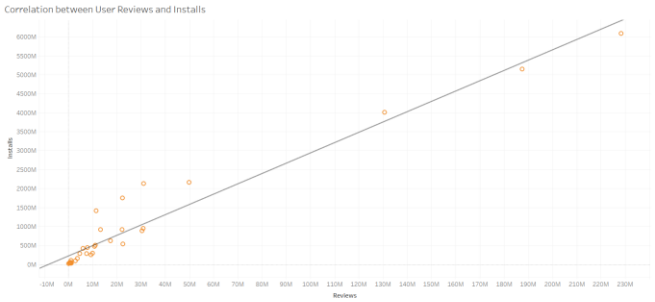

Figure 6. Scatter Plot showing the relationship between Installs and Reviews.

The scatter plot shows a positive correlation between user reviews based on the number of installs made by users. The user reviews are below 10 million records for the number of installs below 500 million. Apart from that, the game has the highest reviews for 228 million with 600 million installs followed by tools and family.

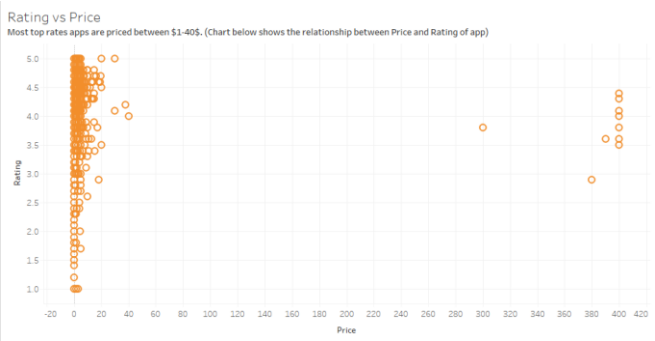

Figure 7. Scatter Plot showing the relationship between Rating and Price.

The scatter plot shows that top-rated apps are mainly by free apps users. While the rating for paid apps is relatively good for those priced better $1 to $40.

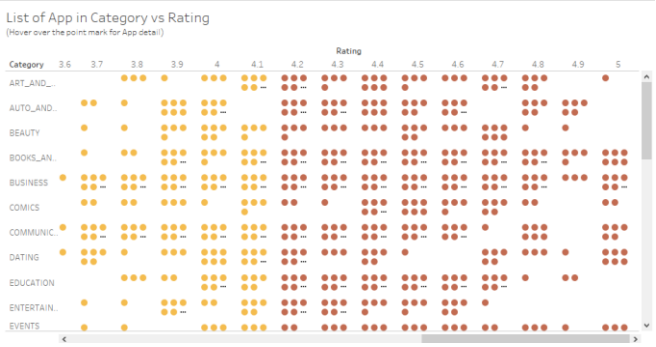

Figure 8. Gantt Chart list all the Apps in each category with Rating.

This is an interactive view. Game highest rating for apps Bingo, Fishing Hunter, Axe Champion. Tools highest rating for alarm clock and age calculator.

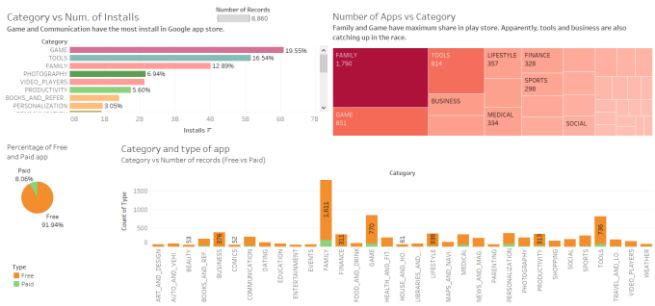

Figure 9. Dashboard Overview

The dashboard has the main features whereby the popularity of categories, number of apps in each category, type of app for each category, and the percentage of the type of app.

7. DISCUSSION AND FUTURE WORK

Tableau software offers a practical solution to messy data which is very difficult to understand. We can make the data live with data visualization. Interactive and meaningful data visualization can make the data live and can tell stories as proverb "stories visible and bring[s] them to life." (Few 2009a, 5).

Around 50% of the world's population which is equivalent to 3.9 billion people, were determined to be online in 2018, and 96% of the world's population lives within the expanse of a mobile network. It was estimated that over 4 billion mobile devices were being used in 2018, with the majority of demographics in mature markets having several devices. It is needless to say that since everything is at the fingertips now, the usage limit of apps and smartphones is bound to increase by multi-folds, as is evident as the time spent using apps globally grew to 50% from 2016 to 2018. From our visualization, we can see that the world population is much focused and fascinated with the apps available mainly in games, tools, and family. Therefore, apps developers should look into options to build emotional intelligence strengthening and robotic formation games. These games will draw the attention of youths between the age of 25 to 35 as it will indirectly build their creativity and this can be treated as their mind relaxer during leisure hours. Population between the age of 25 to 35 will appreciate such games which both diffuse stress and develops their creativity. Some of the higher-level options in the game can be as pay per use within the price rate of $1 to S40 as this is affordable based on the current study. The tools are one of the top-rated categories of apps users. From the visualization, we can clearly understand the most popular and heavily used applications mainly fall into the category of games, family, and tools.

7.1 Strengths

When choosing the color, size, and contrast, it seems to be the correct usage. We have emphasized this as you can see based on the user's view, it falls on the left side first. Our visualization has proven it. Moreover, our data show a variety of ways to visualize it. We didn't stick to one type but multiple designs were used to visualize our data. Furthermore, generating visual graphics aids in showing insights that may have been missed in a normal traditional report. Overall, we could see a trend in all the graphs that have been generated. Finally, we can generate reliable idioms for the created tasks in this study. Thus, it is possible to avoid unnecessary and complex idioms. Other than that, due to the many attributes in the dataset, we can plot many data visualization and visual analytics, thus extracting some of the important features for data analysis.

7.2 Weaknesses

The idioms generated in this study have been based on the instances filter of specific attributes to bring out visual perception to the viewer to fit in the requirement. Besides that, we are not able to do some task analysis due to the complexity of the data and we need to do analysis based on the selected attributes only.

7.3 Limitations

Expanding the analysis required a deep understanding of visualizations idioms and tools to reveal a hidden pattern or insight into the data. The analysis of the app's dataset is a complex system. Thus, we limited our analysis to know the best apps based on the most downloaded and rating features.

7.4 Lessons learned

In the beginning, we wanted to use a boxplot as one of the visualizations. But eventually, we decided not to since boxplot is used to see the data distribution of each category. So we were not sure if it could help to explain the data. Therefore, we decided to go on with what we knew. We realized that we could use interactive visualization but due to time constraints, we would put it in our future work. Furthermore, the creation of idioms depends on the selected task which is by choosing different attributes, will elicit different idioms.

7.5 Future Work

We will look into methods to incorporate the most prominent apps into each visualization to have a clearer insight into each category. If we could have more time we would have used interactive visualization to enhance our graphics which are to be viewed.

Other than that, in the future, we plan to work on percentile, components, sub-components, and color attribute that will be helpful for more measurement and precise analysis of Google play store application. The similarity factor in different applications can be checked, also can make clusters of different attributes relationship between different attributes with the help of doing clustering or aggregation.

8. Conclusion

There are two major types of applications on Google play store are free and paid. These types of applications also have several other categories of applications like games, movies, education, video, and many more. All these applications are present on the Google play store. In this paper, by using a tableau to build a Google play store dataset with many important features we find valuable information which can be very helpful for Play Store Apps developers and as well users.

The smart tools apps can be enhanced not just for the use of home appliances instead such apps can be developed for the use of schools. The apps for attendance, homework submission, students' co-curricular progress can be determined easily with apps that will be highly popular among teachers, students and parents as well. On the other hand, the business unit team should also look into categories that can be improved as they are the potential to be popular such as photography and productivity.

In a nutshell, this visualization is more helpful for game developers in the development phases, also for the users of the game's application for the selection of the game that they want to play, and also for other users who intend to download applications for their uses.

References

[1]"Tableau,"[Online].Available:https://www.tableau.com/learn/articles/data-visualization.

[2]C. Chun-houh, H. Wolfgang & U. Anthony, Handbook of Data Visualization, Springer, 2008.

[3] "Cambridge Dictionary," 20 Oct 2019. [Online].
Available:
https://dictionary.cambridge.org/dictionary/english/garbage-in-garbage-out.

[4] J. Clement, "Number of apps available in leading app stores 2019," [Online].
Available: https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/. [Accessed 20 OCT 2019].

[5] Amir Latif, R. M., Aslam Shah, S. U., Ijaz, F., Abdullah, M. T., Farhan, M., & Karim, A. K. (2019). Data Scraping from Google Play Store and Visualization of its Content for Analytics. *International Conference on Computing, Mathematics and Engineering Technologies – iCoMET*.

[6] Visualization Analysis and Design. Munzner. AK Peters Visualization Series, CRC Press, Nov 2014.