Analysis and Visualization of Google Play Store Apps

Chloe Ngo
chau.ngo@sjsu.edu
015443194

Saritha Podali
saritha.podali@sjsu.edu
015309775

Abinaya Seshadre
abinaya.seshadre@sjsu.edu
015314013

MS, Data Analytics Program,
San José State University

Abstract — One of the most used applications in any Android device is the Google Play Store. Sensing the huge potential and market of Mobile Apps, developers all over the world are contributing to developing a wide range of apps in a multitude of categories, ranging from education to entertainment. With the Playstore's popularity and the rich set of apps in the Playstore, we see potential in how the app data can contribute to making a positive impact on the Android apps market. Through this data visualization research project, we leverage the Google Playstore dataset from Kaggle, Python for data pre-processing, and Tableau for data exploration & visualization. Using the rich capabilities of this tool, we analyzed the Google Play Store apps, where free apps dominate the Android apps market with over 600+ billion downloads and an active user engagement of 5+ billion users of over 1.2 million apps. As a result, we were able to uncover insights like the category of apps with the most installs in a given year and few features of paid apps by comparing and contrasting the dataset attributes using Tableau.

Keywords - Google Play, Google's Android, Data Visualization, Google Apps

I. INTRODUCTION

In 2020, the number of applications downloaded from the Google Play Store is 108.5 which increased nearly 20% from the previous year [1]. For Android users, the Google Play store is a common application where users use the most to download other applications such as social media apps, music apps, games, etc since it has millions of apps of all kinds. As per 2021 statistics, Apple's iOS and Google's Android capture 99% of the Global Market share [2]. Although iOS users are fewer in number compared to Android users, iOS beats Android in terms of revenue generation. We believe the Play Store apps data has huge potential in making app-development businesses successful.

In the dataset provided by Kaggle [3], there are more than 1million data which is 1048576 data and each data contains 24 columns. The data is cleaned and we are going to filter out some columns as they are beneficial for our visualization. We want to visualize and analyze the features of Google

Play store apps by their app name, category, rating, price, and sizes. We believe that analyzing these features would definitely help the app developers know what their customers need and what they actually want while using the apps. This will derive actionable insights to help app developers understand customer demands better thus enabling them to capture the App Market better. Also, we are going to use Tableau to Explore rich interactive data visualization capabilities.

II. RELATED WORK

Shashank, et al. [4] used the feedback from the users to analyze the data from Google Play Store Apps. With the dataset from Kaggle, the authors find out the relationships between numerical ratings, user reviews, and whether the app is free/paid. Prediction is based on the user's rating of the apps together with machine learning methods such as Random Forest, KNN, and K Means Clustering to calculate the accuracy of the methods. The results show that KNN gives out the highest accuracy and predicts that with more than 100,000 installs, it will be successful on Google Play Store Apps.

M.Armir Latif, et al. [5] wanted to build a Google play store based on Google-play-scraper. The authors scrape all the categories in the game including free and paid apps. Also, the authors use CIRCOS and histogram to visualize the relationship between the free/paid apps with the user's rating of each category. The main focus of this visualization is to help the game developers as well as the users whose interest in gaming apps.

Rimsha Maredia [6] also used the dataset from Kaggle to visualize the data and predict the popularity of the apps on the Google Play Store. The author visualized the number of apps that are installed by the users separated into individual categories and found the average ratings. In addition, the author has applied machine learning such as decision tree, KNN, Naive Bayes, and Logistic Regression to compare the accuracy of each method.

III. METHODS

A. Dataset

For this project, we made use of the Google Play Store Apps dataset found in Kaggle.com. It has around 2M+ records covering about 600k+ Google Apps. It also has many useful columns, of which the key columns include the App name, Category, Installs, Price, Size, Rating, and so on. Below is a brief description of the key columns.

DATA VARIABLES

Variables	*Description*
App Name	Gives the unique names of the Apps in the Play Store.
Category	Denotes the category to which the apps belong. Example categories are Education, Social, Entertainment, etc.
Installs	Tells how many times the app has been installed or

	implicitly, the number of users for an app
Size	Gives the size of each app in megabytes
Rating	Gives the average rating for each app
Rating Count	Denotes how many times or in other words, by how many users the app has been rated.
Free/ Paid	Tells if the app can be installed and used without any cost to the user.

Table 1: Data variables and their description

B. Data Cleaning and Preparation

Data cleaning & preparation is a critical step in any Data Analysis & Visualization project. Python programming language was used to do the data preparation & cleaning to generate a cleaner dataset for Tableau. Missing values & duplicate records are common problems with unclean data. It is important to handle them, if not, it can significantly alter the outcome of the data analysis. Besides unnecessary columns such as *App developer, security policy, App ID, editor's choice*, etc were removed as they do not add any value to the data visualization problem our project deals with. Date format inconsistencies can make the analysis difficult, hence, ensuring the date fields such as *Released* & *Last Updated*, adhere to the format supported by the visualization tool is very important. Categorical columns often have values that imply the same, for example, application categories *Music* and *Music & Audio* are the same, but they exist as two separate categories in the dataset. So, it is important to ensure the values are accurately grouped. The last data preparation step that our dataset required was adjusting the scale of the column values - the field 'Size', had values ranging from kilobytes to gigabytes. In order to do a reasonable analysis, all the values were converted into *megabytes* for uniformity.

The clean dataset which is of .csv format was then imported into Tableau for visual exploratory data analysis.

C. Tableau Workflow

Tableau supports the usage of .csv files as the data source making it an ideal visualization tool for simple to complex data analysis. Upon importing the clean data into Tableau Desktop, we can see that Tableau conveniently does the job of separating dimensions from measures and also assigning appropriate data types. Each data view we would want can be created in individual worksheets. This allows visualizing data with respect to individual questions we would want our data to answer. For this purpose, we leverage Tableau features such as filters, charts & parameters. Once we have all the worksheets prepared, we can put them together in a single place, the dashboard, to make a logical sense and build a narrative around the data. The scope of this project is limited to building a single dashboard. The final dashboard is then

exported as a .twbx file that can be easily opened using a Tableau reader for viewing or a Tableau desktop for other developers to work on it further. For public visibility, we intend to publish our dashboard to Tableau Public. Fig. 1 illustrates the workflow mentioned in our visualization project.

Fig. 1: Tableau Workflow

IV. RESULTS

The objective of our visualization project was to do an exploratory analysis of the Google play store apps data to understand the data patterns, derive appropriate information to present it in a useful interactive dashboard. The final outcome of the project is as shown in Fig. 2, built for the Generic Desktop display option on Tableau.

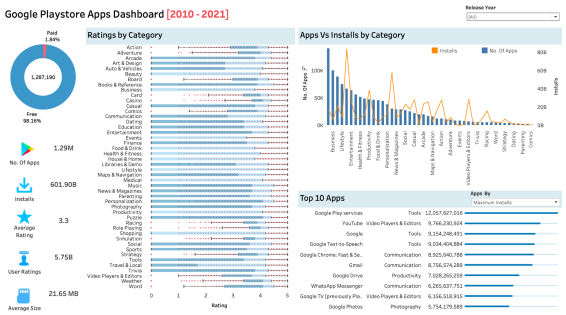

Fig. 2: Dashboard for Google PlayStore Apps Analysis and Visualization.

Fig. 2 shows that all the visualization elements are captured in a single interactive dashboard. The dashboard has global filters, such as *Release Year* that allow users to visualize data based on years ranging from 2010 - 2021. The key metrics such as no. of applications, installs, ratings, user ratings & size of the apps are available with image representation for easier interpretation. These metrics can be availed using the aforementioned global filter. Since the dataset has a huge number of applications belonging to different categories, we introduced parameterized filters to allow users to drill down into the data by various factors such as installs, price, size, etc. The parameterized filter *Apps/Categories By* allows users to find the top 10 applications & categories by each of the respective contributing factors. Besides the dashboard has other visualizations such as a donut chart and dual-axis line chart. Appropriate scaling & legends are applied.

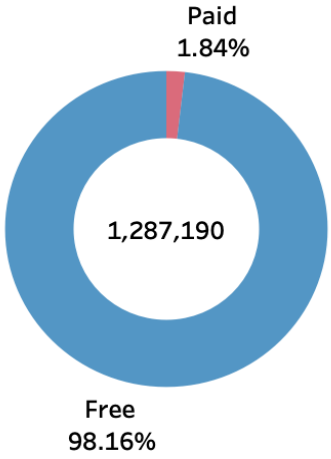

Fig. 3: Free Vs Paid apps

All the apps in the Google Play store are either free or paid apps. Fig. 3 shows that the percentage of free apps is very high when compared to paid apps. Paid apps contribute to only approx. 1.84% and approx. 98.16% are free apps of the total number of applications available in the play store.

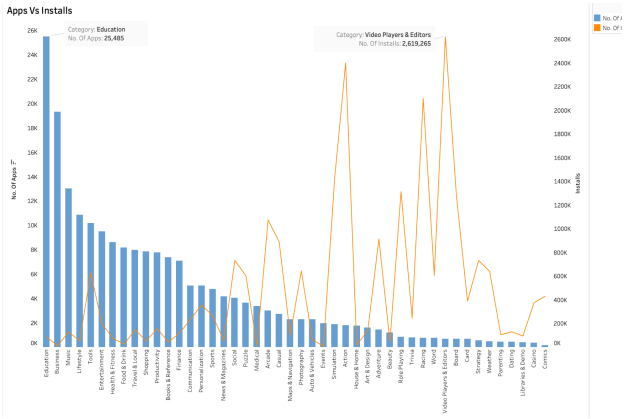

Fig. 4: Number of applications versus the number of installs

Fig. 4 suggests that over the years, although the highest number of apps are available in the Education Category, the trend suggests that the highest number of app installations are in the Video Players & Editors category. The growing popularity of YouTube, Instagram & TikTok is an outstanding example of this.

Fig. 5: Top 10 Apps filtered by Maximum Installs from 2010-2021.

Fig. 5 illustrates the top 10 apps filtered by maximum installs. Bar charts are used to compare the installs between the apps and the categories. Visualizing the top 10 Apps, Google Play services appeared to be the app that has the most which are 12,057,627,016 installs. Youtube comes in second place with 9,766,230,924 which is around 3 million difference between the first and the second place. Google TV and Google Photos are the apps that have the lowest among the top 10 apps during all year with 6,156,518,915 and 5,754,179,589 installs, respectively.

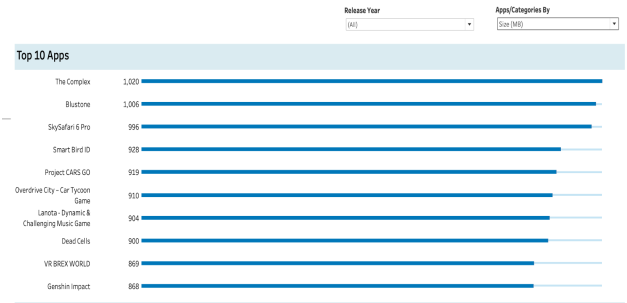

Fig. 6: Top 10 Apps filtered by Size (MB) during 2010-2021.

Fig. 6 illustrates the top 10 apps filtered by size. Our team has used a bar graph to show the difference in size for each app as well as the categories. All the sizes measured are in megabytes(MB). For the size of an app, as it is shown from the graph, The Complex has the biggest size which is 1020 MB and followed by Blustone app which size is not much different, with 1,006 MB. VR Brex World and Genshin Impact are the two apps with the lowest size among the top 10 of all years. They consumed around 869MB and 868 MB, respectively. So in general, The Complex and the Genshin Impact has around 150MB difference between the largest app size and the smallest app size of the top 10.

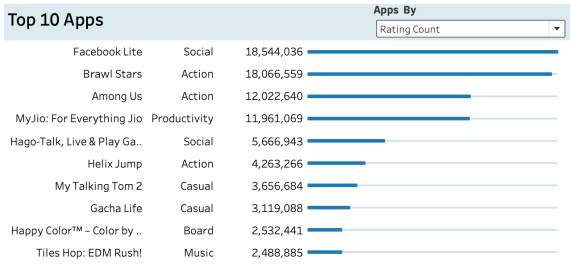

Fig. 7: Top 10 Apps filtered by Rating Count during 2010-2021.

Fig. 7 illustrates the top 10 apps filtered by rating count during the years 2010-2021. In this figure, we also use a bar graph for a better understanding of which app and category have the highest rating. For apps, Facebook Lite is reported to be the app with the highest rating count of all the apps with 18,544,036, and Tiles Hop is the app with the lowest rating count in the top 10 which is a Music app with 2,488,885. Since communication and social apps are quite popular for smartphone users of all ages, it is not a surprise when the rating count of these apps is the highest among all.

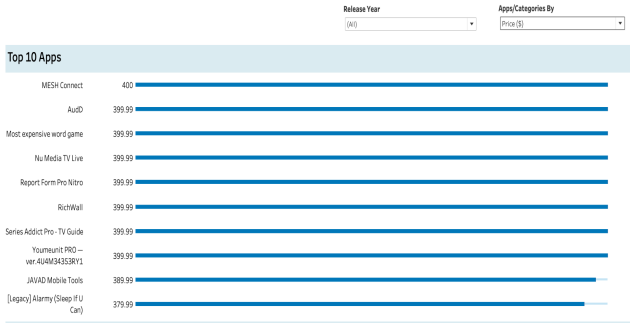

Fig 8: Top 10 apps filtered by Price ($).

Besides the free apps, there are some apps that require users to pay in order to use. Fig. 8 illustrates the top 10 apps that are filtered by Price. The most expensive app of all years is MESH Connect which costs $400. AudD came in second place with $399.9 which is roughly the same price as MESH Connect and as well as Word game, Nu Media TV Live, Report Form Pro Nito, RichWall, Series Addict Pro, and Youmeunit Pro, which all cost the same price. Alarmy comes in last place with $379.99.

Using the box-plot visualization technique as seen in Fig. 9, we were able to visually see the average ratings of each of the application categories. It is very apparent from the dataset that there are outliers in the dataset that can skew the app ratings. However, using the box-plot allows us to keep the outliers out of the computation of average ratings. Overall, the play store data has an average rating of approx. 3.3.

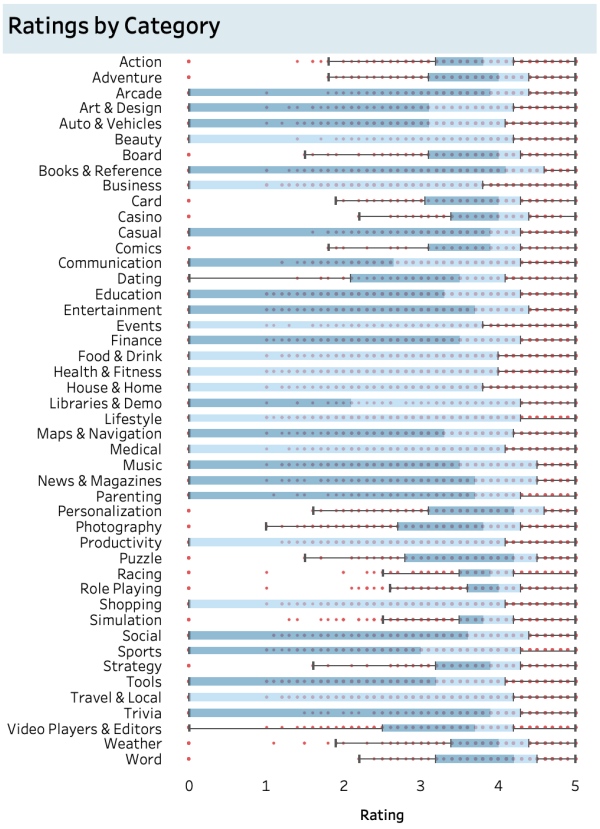

Fig 9: Box plot showing ratings by category

V. DISCUSSION

After coming up with our individual worksheets and charts in Tableau, we combined them together in a dashboard. Looking at the dashboard, with all charts grouped together, we were able to spot a few shortcomings and felt it could be refined more as discussions on the entirety of the visuals going on. We also shared our dashboard with a few of our friends and family and got feedback. Below is the summary of the opinions, problems, discussion, and changes we made that improved our dashboard.

Firstly, we had too many textual components, which produced a mundane visual and made information distinction difficult, so we resorted to bringing in icons (Fig. 10) to represent the attributes of our apps. Then, people were curious to know how the data changed over the years, so we brought in global date filters to show the trend.

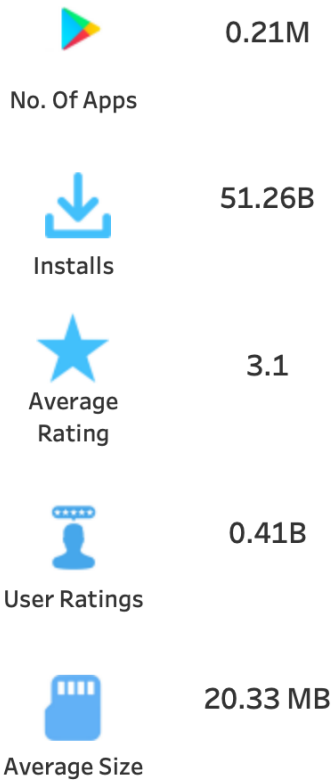

Fig. 10: Icon-based metric summary

Once we improved our dashboard, we showed our audience and they were clearly able to capture more significant information. They were able to capture information like the number of installs and the average rating for apps for any selected year. They came to know the top 10 apps and the top categories of apps that had the maximum installs or maximum ratings, depending on the option they selected on the drop-down filter (Fig. 11).

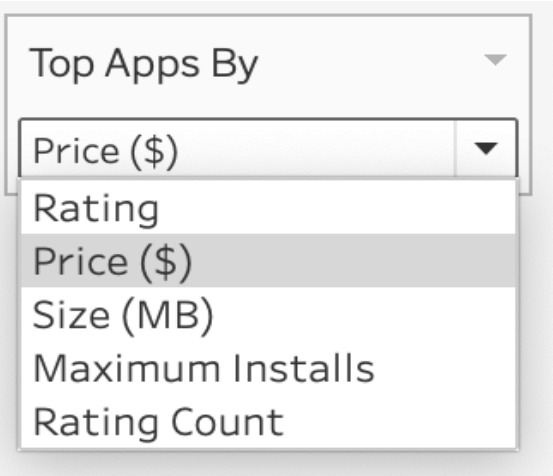

Fig. 11: Parameterized Filter

The audience and we were also convinced that following important principles like the Data Ink ratio where noise should not exceed useful data and Gestalt principles like having continuity and proximity and grouping similar elements together in our design enhances readability and clarity. Lastly, apt and concise titles, labels, and scales are required for the audience to make sense of the visualizations.

We conclude the dataset has great potential to allow analysts to work on the business values and make necessary adjustments in order to make a positive impact on the company's business and customer value. Dataset required only a few data cleaning & preparation steps which makes it an ideal and much cleaner dataset compared to most others which usually require huge amounts of data cleaning, pre-processing & preparation. The dataset is appropriate for data visualization analysis and ML projects.

VI. FUTURE WORK

Apart from the analysis and visualizations with the Play Store app data in this project, more analysis with respect to other features such as the release date, developer ID can be used for uncovering even more correlations with the demand and consumption of the Play Store Apps. ML models can be trained and developed to predict new ratings or update future ratings of the apps. ML/AI algorithms can also be implemented to detect policy violations or other security concerns such as spammy developers or malicious content, etc.

VII. WHAT DID YOU LEARN?

We were able to learn important lessons pertaining to visualization and Tableau from executing this project. First of all, we learned that clean data is very important for Tableau. Second, the Speed of data load and command execution in the Tableau free version varied between system configurations. Only one of the team members' Tableau loaded slower than the rest, showing Tableau's great capability with 1 million+ rows of data.

Although filtering is a great feature in interactive dashboards, having too many features can degrade the performance of the dashboard. So we also acknowledge that updating the dataset locally and refreshing the data store in Tableau automatically updates the visualizations and dashboards.

Based on this project, we have learned that data types can be changed at the Tableau data store. Following the Data Ink

ratio and Gestalt principles of visualization play very important roles in making the important information visible, accessible, and easily comprehensible in a dashboard. The charts need to be related, continuous, and in proximity as learned in the class. This helped us come up with a theme that helped us separate the necessary and unnecessary elements. We also tried to reduce data redundancy by excluding unnecessary repetition of fields and labels. In this way, we were able to design a clean dashboard with keeping the information that matters clearly visible to the audience.

VIII. REFERENCES

[1] Published by Statista Research Department, and Sep 24. "Google Play Annual App Downloads 2020." *Statista*, 24 Sept.2021, www.statista.com/statistics/734332/google-play-app-installs-per-year/.

[2] Published by S. O'Dea, and Jun 29. "Mobile OS Market Share 2021." *Statista*, 29 June 2021, www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/.

[3] Prakash, Gautham. "Google Play Store Apps." *Kaggle*, 17 June 2021, www.kaggle.com/gauthamp10/google-playstore-apps.

[4] Shashank, S., & Naidu, B. (2020, December). *Google play store apps- Data Analysis and ratings prediction*. Retrieved October 7, 2021, from https://www.irjet.net/archives/V7/i12/IRJET-V7I1248.pdf.

[5] R. M. Amir Latif, M. Talha Abdullah, S. U. Aslam Shah, M. Farhan, F. Ijaz, and A. Karim, "Data Scraping from Google Play Store and Visualization of its Content for Analytics," 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2019, pp.1-8, doi: 10.1109/ICOMET.2019.8673523.

[6] Maredia, Rimsha. *(PDF) Analysis of Google Play Store Data Set and Predict ...* www.researchgate.net/publication/343769728_Analysis_of_Google_Play_Store_Data_set_and_predict_the_popularity_of_an_app_on_Google_Play_Store.

IX. APPENDIX

Raw data was downloaded from Kaggle and cleaned using python which was then imported into Tableau and used as the data source for this visualization project. Fig. 12 shows the data imported into Tableau to build the workbook.

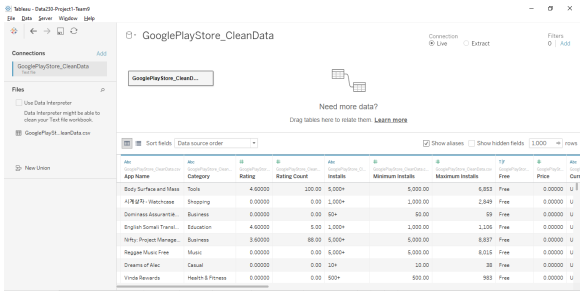

Fig. 12: Data source