Data Analytics
UE20CS312

Mini Project
Literature Review & Initial Solution Approach

Sakshi Hulageri - PES2UG20CS300
Saakshi H Srinivasan - PES2UG20CS290
Sanjana Pai Kasturi - PES2UG20CS309

Dataset chosen: Google Play Store Apps
(https://www.kaggle.com/datasets/lava18/google-play-store-apps)

EDA and Visualization:

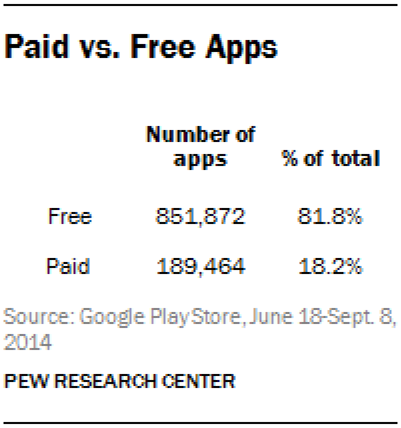
https://github.com/Data-Analytics-Team3/DA_Literature_review

- What have others done to solve this problem? What other approaches can we explore on this data set?
  Or
- How have others solved a similar problem? Can we apply any of those solution strategies to the problem we have selected?

At the time of the data collection according to the paper Chapter 2: An Analysis of Apps in the Google Play Store , the Google Play Store broke apps down into 41 general categories. Education apps were the most common individual category, comprising 8% of the total number of apps available for download.

Overall, eight categories of apps (Education, Entertainment, Personalization, Tools, Lifestyle, Books and Reference, Business, and Travel & Local) comprised more than half of the apps available for download (53.58% in total).

Music apps were the least prevalent category, comprising just 668 apps — or 0.06% of the more than 1 million total apps in the Store. When collecting this app data, Pew Research Center used the categories in the Google Play Store and conducted no additional categorization of the apps in the dataset.



Paid vs. Free Apps

	Number of apps	% of total
Free	851,872	81.8%
Paid	189,464	18.2%

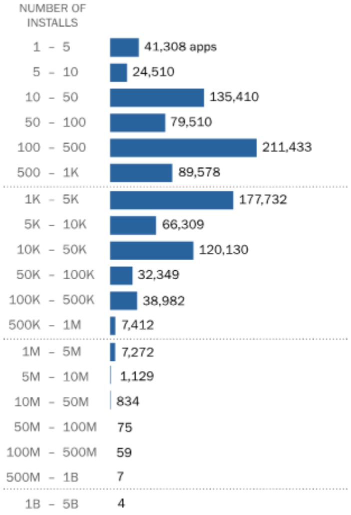
Source: Google Play Store, June 18-Sept. 8, 2014

PEW RESEARCH CENTER

The majority of apps in the Google Play Store (82%) were free to download at the time of the data collection. Most, but not all, apps that are free to download were supported by advertising. On average, free apps ask for two more permissions than paid apps (Six permissions vs. four permissions.)

Android Apps Have a Long Tail by Number of Installs

# of apps by the # of times they have been installed



Source: Source: Google Play Store, June 18-Sept 8, 2014
Note: 7,293 apps did not have install information.

PEW RESEARCH CENTER

The Google Play Store contained more than 1 million apps, but the overwhelming majority of these apps had been installed by only a small number of users. Close to half (47%) of all apps available had been installed fewer than 500 times, and more than 90% had been installed fewer than 50,000 times. On the other end of the spectrum, a relatively small number of apps had been installed by vast numbers of users.

We aim to establish similar relationships between the attributes in our dataset, showing which apps are rated the highest, their popularity and improving knowledge of the apps for the users.

In another paper, https://www.researchgate.net/publication/343769728_Analysis_of_Google_Play_Store_Data_set_and_predict_the_popularity_of_an_app_on_Google_Play_Store/link/5f3e8d66a6fdcccc43d8e3cb/download, classification models like KNN, Logistic Regression, Gaussian Naive Bayes, and Decision Trees have been used.

We can also employ these methods for our model, preferably, one that gives high accuracy such as logistic regression.

Refine your Problem Statement:

- What is the specific problem we are going to solve?
  We are providing a statistical model of apps, their ratings and reviews to the users.

- What are the questions we are going to attempt to answer?
  Which is the highest rated app?
  What people are mostly buying?
  Number of reviews received on which type of apps

- What are the challenges with this data set (based on the initial exploratory analysis + coarse solution approach (trying library functions, etc., to build a simple model)
  Combination of many data types and many rows(10000)
  There are null and missing values
  There is 1 outlier

- What solution approaches would be reasonable to attempt?
  Outlier detection and removal, data imputation and manipulation using mean, median and mode where mode values were used for categorical values.

- How is my solution approach different from what is already out there?
  We have used the Fillna() function to fill null values with mode values .If its bimodal we take the first value.

- What is the use of solving this problem?
  The users get an idea about which app they should be installing based on their requirements and also get a detailed description of every app.

- Literature Survey Report

As data availability, completeness and accuracy is a big issue in the mobile app market, the work in this area is still very limited. Google Play store ranks each app that is published in the store. The overall app ranking system of Google (Fernandez 2013) uses very complex infrastructure to rank apps. This is the reason google app ranking system outranks other apps stores and is very successful. According to the researchers, Google Play is more of a superstar market due to its popular products and apps. Google play has a nice and clean adoption system that ranks apps (Zhong and Michahelles 2013). Also, these ranking systems are very complex and rank the apps in a very good manner, there are many cases where users scroll the pages to the bottom and find the apps they like or find an app with a better rating. Since the introduction of mobile app stores, there are a few researchers who work in the area of mobile app rating prediction and variable importance that helps in finding a correlation between app ratings and other factors. Researchers have worked in different domains to find out the ways to get better ratings in apps. Some researchers focused on user’s reviews, some researchers targeted the app's attributes and features, and some researchers worked in the field of better software engineering practices. However, all of these domains are important at their places, but usually app attributes are analyzed by the researchers to find a relationship between app rating and its attributes. In this section, we discuss some of the contemporary research works in this area. Tian et al. (2015) performed a case study using statistical analysis to rank the different factors of apps that affect the app ratings, the size of an app, promotional images and target sdk of an app are the most influential factors of high-rated apps. Similarly, Finkelstein et al. (2017) investigates the relationship between price, rating and popularity in the blackberry app store and their findings show that there is a strong correlation between customer ratings and popularity. Researchers performed a detailed analysis of apps from Android and Apple apps and performed a quantitative analysis of app attributes and their effects on the apps in different app stores (Ali et al., 2017). Moreover, Liang et al. (2017) used feature-oriented matrix factorization to predict the mobile application ratings. Researchers uncover factors that influence the app rankings for apple app stores and proposed a model that predicts the ratings for different apps. They considered a number of variables in their model, including package size, app release date, category popularity, etc. to find the importance of these factors (Picoto et al. 2019). Khalid et al. (2016) performed an analysis of finding the relationship between app ratings and static analysis warnings. According to their findings, the developers can use static analysis tools to identify bugs. Similarly, researchers used mobile app ratings for the app recommender system, expert systems and knowledge based system for different domains. Researchers also proposed models to rank the risks of android apps using different machine learning models (Peng et al. 2012). However, the area of app variable importance is very limited and there is a gap in this field. Also, at one hand, developers and companies try their best to make apps to gain better rankings. On the other hand, researchers also identify ranking frauds in the mobile app market carried out by the companies and developers to gain better rankings (Zhu et al. 2015) Although some researchers suggest that app rating is not considered important or is variable like Liu et al. (2014) performed a detailed analysis of Google play store. According to their findings the review ratings have lower impact in case of free apps. In another research Martin et al. (2016) performed a detailed analysis of app releases by developers. According to their findings, 33% of such releases caused a significant amount of change in user ratings. Karagkiozidou, Makrina, et al. conducted a research study that helps the developers using the proper keywords and other primary things to gain better rankings of the apps using App Store Optimization (ASO) (Karagkiozidou et al. 2019). Similarly, McIlroy et al. (2017) performed an analysis of google play app ratings when the company responds to those ratings. The results show that users change their ratings 38% of the time following a response. However, sometimes, for the user's own satisfaction, sometimes by the requirements and the threshold of recommender systems and expert systems, it is beneficial for an app to have high ratings. Such apps usually gain more downloads and are more attractive to the users.

References:

Zhong N, Michahelles F (2013) Google Play is Not a Long Tail Market: An Empirical Analysis of App Adoption on the Google Play App Market, in: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC ‘13. ACM, New York, pp. 499–504. 10.1145/2480362.2480460

McIlroy S, Shang W, Ali N, Hassan AE (2017) Is It Worth Responding to Reviews? Studying the Top Free Apps in Google Play. IEEE Softw 34:64–71. https://doi.org/10.1109/MS.2015.149

Tian Y, Nagappan M, Lo D, Hassan AE (2015) What are the characteristics of high-rated apps? A case study on free Android Applications, in: 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME). Presented at the 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 301–310. 10.1109/ICSM.2015.7332476

Zhu H, Xiong H, Ge Y, Chen E (2015) Discovery of Ranking Fraud for Mobile Apps. IEEE Trans Knowl Data Eng 27:74–87. https://doi. org/10.1109/TKDE.2014.2320733

Karagkiozidou M, Ziakis C, Vlachopoulou M, Kyrkoudis T (2019) App Store Optimization Factors for Effective Mobile App Ranking. In: Kavoura A, Kefallonitis E, Giovanis A (eds) Strategic Innovative Marketing and Tourism, Springer Proceedings in Business and Economics. Springer International Publishing, New York, pp 479–486

Khalid H, Nagappan M, Hassan AE (2016) Examining the Relationship between FindBugs Warnings and App Ratings. IEEE Softw 33:34– 39. https://doi.org/10.1109/MS.2015.29 Liang T, Chen L, Ying X, Yu PS, Wu J, Zheng Z (2017) Mobile Application Rating Prediction via Feature-Oriented Matrix Factorization, in: 2017 IEEE International Conference on Web Services (ICWS). Presented at the 2017 IEEE International Conference on Web Services (ICWS), pp. 261–268. 10.1109/

ICWS.2017.41 Liu CZ, Au YA, Choi HS (2014) Effects of Freemium Strategy in the Mobile App Market: An Empirical Study of Google Play. J Manag Inf Syst 31:326–354. https://doi.org/10.1080/07421222.2014. 995564 Martin W, Sarro F, Harman M (2016) Causal Impact Analysis for App Releases in Google Play, in: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2016. ACM, New York, pp. 435–446. 10.1145/ 2950290.2950320