Investigation of Google Play Store Application Downloading in 2017

Ziyue(Hannah) Zhang

Denison University **Department of Mathematics and Computer Science**

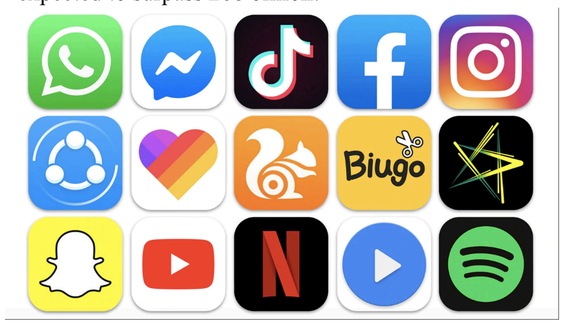
Recently, a new shopping method started to appear: application shopping, as the smart phones become a part of people's lives. People would like to spend increasingly amount of money and time on their device. They start to replied to apps providing services and assisting their lives.

Here we used a dataset which collecting data from Google Play store, one of the largest apps stores. It recorded all the apps' basic information and downloads in 2017. By using this dataset, we want to analyze people's shopping behaviors among various apps, and conclude some patterns behind.

OBJECTIVES

Nowadays, most people have their own smart phone. People used various apps on the phone to assist their work and make their life easier.

In fact, by 2017, the number of downloaded apps is expected to surpass 268 billion!



One unexpected fact is some studies showed that more than 50% of the apps that are downloaded are deleted after use once. In other words, people rarely use for almost half of their apps. Why and How that's happened?

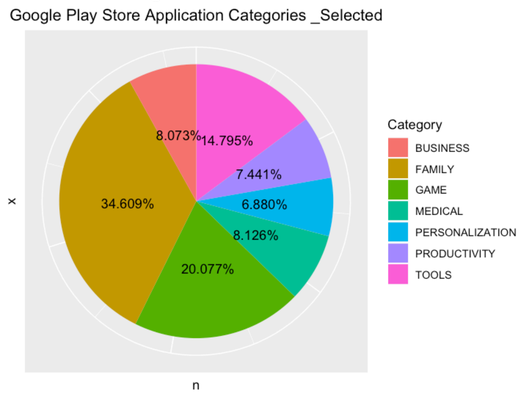
Generally speaking, suitableness and usefulness are the most essential factors when customers shopped. However, Under today's App markets, there are other factors that may affect people's downloading behavior, like other users' feedback and some objective factors, like size.

The main purpose of this project is to analyze people's behavior about downloading. **How other people's opinions affect new consumers' preference of downloading an app? Does objective factors like size affect people's purchasing decision? In positively or negatively way?**

METHODS

We obtained data from Google Play Store, which can be accessed in kaggle, last updated in two years ago. About 10k entries about application downloading in 2017 reflected in the data.

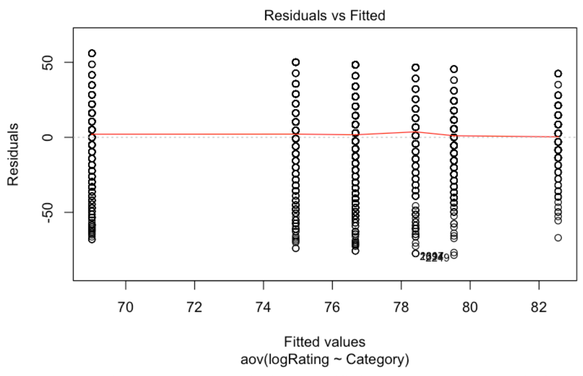
We edited down the original dataset from 10841 to 3863 by choosing to focus on the main categories only, as there were 34 categories in total. We selected seven categories, including business, family, game, personalization, medical, productivity, and tools, because of the seven most common categories.



For the first step of the project, we used the ANOVA test to explore mean differences of these categories. Then, we chose three possible factors: Rating, Reviews, and Size, that may affect our dependent variable installs. The first two mainly explore how other people's feeling impacting new consumers behaviors, and the last one tested whether objective factor size influence people's preference of downloading.

In the original dataset, rating and reviews are numerical variables, but the size are the character variable. Then, we converted the size to the numerical variable too by deleting the unit M, that will be helpful for our analysis further. Hence, we applied generalized logistic regression for Poisson distribution, chi-squared test, and many other statistical approaches to determine whether these correlations are significant.

For almost all models the residuals are not normally distributed, so we did the data transformation using logarithm, square roots, and cube operations in terms of making both tests satisfy assumptions.



Next, we used Pearson's correlation coefficient to measure a relationship between each models. Then, we used a multiple regression model to try to find the best predicting model, and get our final conclusions.

RESULTS

By conducting several statistical tests: including not limited to ANOVA, generalized logistic regression, chi-square test, Pearson's correlation test. We concluded those results.

RATING

The result is there are significant evidence showing that a higher rating means a higher installs.

Firstly, we tested the assumptions for the original dataset and found that skewed to the right a lot. Then we did the data transformation: taking a cube operation to rating data, the dataset now satisfied our assumptions. It is a reliable generalized logistic regression. This test had an estimated value 1.699e-02. We have a value of 2.2051e+11 on 4309 degrees of freedom. Including the variable, it decreased to 2.1416e+11 on 4308 degrees of freedom, a significant reduction in deviance. That implied this factor rating was appropriate to choose that may affect the installs. Lastly, we took a Pearson's correlation test, the correlation constant is about 6.922%. It is not perfectly correlated, but it still had some impacts.

REVIEWS

More reviews will lead to a higher downloading numbers for these seven categories of apps in Google Play Store.

Similar to the rating factor, it was a left skewed data too. We took the logarithm to all the data to do the data transformation. Since, the reviews data here was over dispersion, so it is hard to apply Poisson distribution. Therefore, we chose negative binomial distribution instead, using glm.nb function. The results showed that p-value is smaller than 0.05, and the estimate coefficient was 2.744e-09. That showed more reviews will lead more installs. The residual deviance was 4721.5, which was smaller than the null deviance, which demonstrates that this model was significant.

SIZES

The conclusion of the size analysis is on average a larger size of the apps will result in more installations.

Size was another objective factor that may affect people's downloading behavior. After doing the data transformation by taking the square root, we applied a generalized logistic regression model. Luckily, the result was also significant, with a positive correlation about 7.230e-02, and fell in deviance. We used Pearson's correlation test, the correlation coefficient was 9.28%, which was the most significant number among all these three factors. That illustrated that objective factors affected more than the subjective personal ideas in this project dataset.

Multiple Logistic Regression

Combining all the information above, we concluded that ratings, reviews, and sizes are significant factors affecting installs.

This multiple generalized logistic model better fit, compared with individual comparison making above.

First, we compared the deviance residuals, which are a measure of model fit. The deviance residuals for individual cases was 2.2051e+11 and after the model fit the deviance residuals fell to 1.4279e+11, means this model fit well. The difference between null deviance and residual deviance was larger than the previous model. Using this multiple factors model worked better. Second, observing p-values for three factors, all of them were smaller than 0.05, which demonstrated that all of them are significant in this model and were not redundant.

CONCLUSION

In this project, after selecting the categories we want to explore, we made the hypothesis that these factors, ratings of the apps, other's reviews, and the sizes of the app, might correlate to the number of installs. The results we received by doing the data analysis were significant, all of these had impacts to the installs in different extents.

1. **Higher ratings led to more installs.**
2. **More reviews caused more installs.**
3. **Larger sizes of the apps were more likely to get more installs.**

Adding all these three factors, the **Multiple Regression Model was more significant** represented the selected dataset. Thus, other people's opinions will influence new consumers' decisions, and objective factors will affect too.

REFERENCE

Akolawala, T. (2019, May 17). TikTok Leads App Store Downloads for Fifth Quarter in a Row: Sensor Tower. Retrieved November 17, 2020, from https://gadgets.ndtv.com/apps/news/whatsapp-most-downloaded-app-worldwide-tiktok-leads-ios-app-store-2039032

Gupta, L. (2019, February 03). Google Play Store Apps. Retrieved November 17, 2020, from https://www.kaggle.com/lava18/google-play-store-apps

Iowa State University. (2015, September 30). Mobile apps and online reviews influence consumer behavior. ScienceDaily. Retrieved November 15, 2020 from www.sciencedaily.com/releases/2015/09/150930092506.htm

Google Play

ACKNOWLEDGMENT & CONTACT INFORMATION

Thank you to the Denison University Math and Computer Science department Math 220 class final project and my professor Dr. Zhe Wang for opportunity to sever as the research for this project.

For more information:
Contact: Ziyue(Hannah) Zhang at zhang_z2@denison.edu

DENISON