# Android Play store Data Analysis

*Bikkavolu Prasanthi - B20CS010*

## INTRODUCTION

As the name implies, Android Play Store Data Analysis analyses data that has been taken from the Android Google Play store. We are aware that the Google Play store has more than 3.04 million applications. With the use of several python libraries, I plan to study a variety of apps from the Google Play Store. I have studied 30 apps under the Finance category. From there, I have analysed the kind of bugs that arise within this category. The results, conclusions and analysis of the findings have been put forth in this report.

## METHODOLOGY

### *Overview*

I obtained the data required from the Play store by ***data scrapping*** through google-play-scraper, a python library and web scraper, google chrome extension. I have done ***exploratory data analysis*** (EDA) on these apps. Then I have done ***topic modelling*** (LDA model) on these reviews and divided the reviews into 6 major topics. Based on these topics, I split them into technical and non-technical reviews. I have then done bug analysis for all the technical reviews. A similar process has been followed for all the apps. Then using **CWE** I categorised the issues into errors.

### *Data Scrapping*

The act of importing data from a website into a spreadsheet or locally stored file on your computer is known as data scraping often referred to as web scraping. It is a method that has the ability to autonomously extract data from websites, databases, business applications, or legacy systems. It's one of the best methods for obtaining information on the internet and, in certain circumstances, for sending that information to another website. I have used google-play-scraper, a python library and a web scraper, a google chrome extension.

***Datasets***: Thus I got two datasets, one contains reviews of 30 apps with scores and the other has general data like overall rating, downloads, the total number of reviews, and rated for ages.

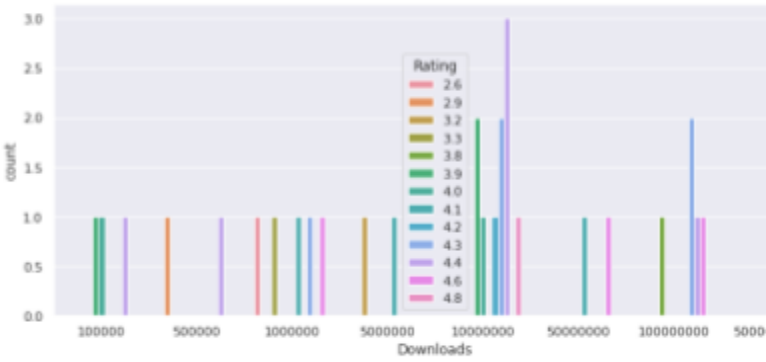
Finance apps, Reviews

### *Exploratory Data Analysis*

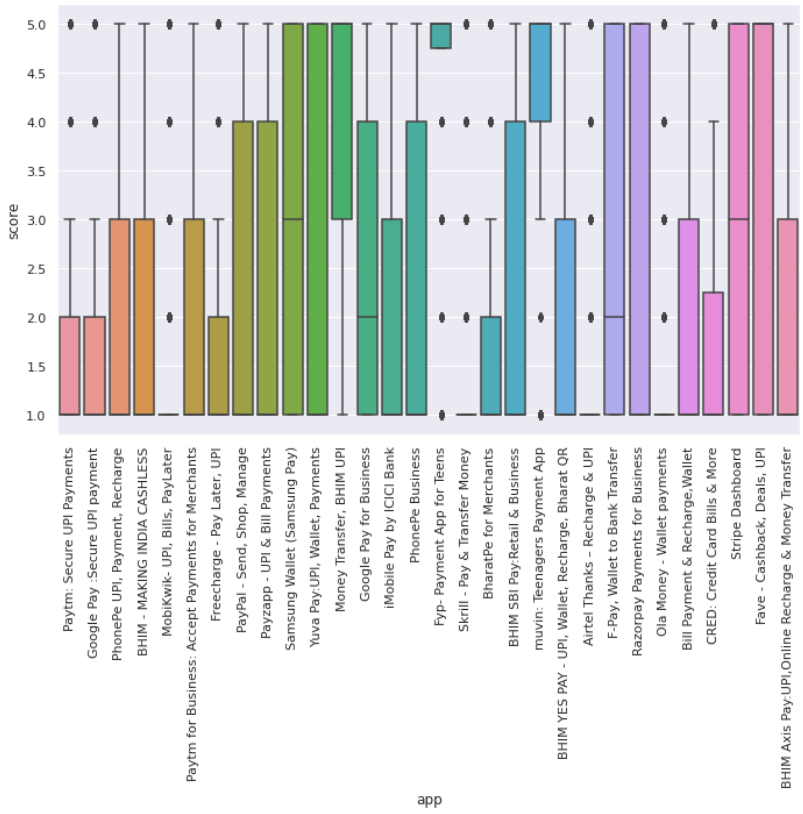
For exploratory data analysis after loading the data in a collab file, first, some preprocessing is done to the data where in some columns some special characters are present, they are removed and then those columns are converted to int or float data types. From the correlation heatmap below, we can say that Reviews and Downloads are somewhat correlated.

**Fig 1.1**: Correlation Heatmap for the dataset


**Fig 1.2**: Number of Apps with Downloads and rating




I have also plotted the box plot for apps with the score distribution across the reviews.




### *Topic Modeling the Reviews:*

Topic modelling is recognizing the words from the topics present in the document or the corpus of data. This is useful because extracting the words from a document takes more time and is much more complex than extracting them from topics present in the document. This looks simpler than processing the entire document

and this is how topic modelling has come up to solve the problem and also visualise things better. After importing the required libraries and the data as a data frame, some preprocessing must be done to the data to make out modelling technique easier.

### *Text Preprocessing*

The following steps are performed during the text preprocessing.

*Tokenization* - Splitting the text into sentences and sentences into words, lowercasing the words and removing punctuation marks.

Words smaller than size 3 are removed.

Stop Words are removed.

*Lemmatization* - Words present in the third person are converted to the first person and words in the future tense and past tense are converted into the present tense.

*Stemming* - words are converted to their root forms.

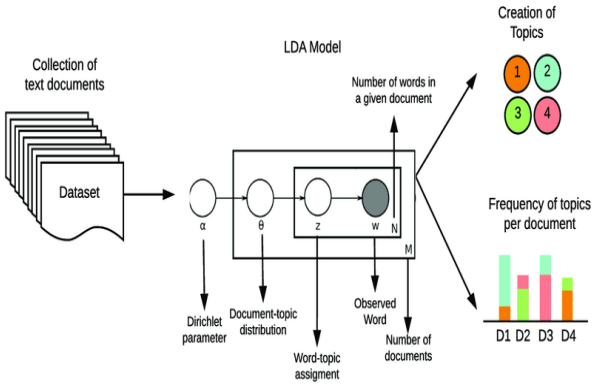
A function is being defined to do all the above preprocessing steps for the review data of each and every app when it is being called.

These are the two functions that are being defined to preprocess the app's review data.

### *Latent Dirichlet Allocation- LDA*

LDA is a topic model that generates topics based on the word frequency from a set of documents.

LDA generates probabilities for the words using which the topics are formed and eventually the topics are classified into documents. Latent Dirichlet Allocation (LDA) does two tasks: it finds the topics from the corpus, and at the same time, assigns these topics to the document present within the same corpus.




### *Bug Analysis*

I analysed the issues identified from the reviews after Topic Modelling using ***Common Weakness Enumeration.***

## RESULTS

After topic modelling, from the top 6 topics, issues have been identified.For the entire category:

1. Negative Review: Support related issues.(non-technical)
2. Positive Review: Easy transactions and User friendly interface.(technical)
3. Negative Review: Support related issues.(non-technical)
4. Positive Review: Many features(technical)
5. Negative Review: Update issues(technical)
6. Negative Revew: Login and Authentication issues.(technical)

For each app:

| Apps | Payment Issues (tech) | Support Issues (non-tech) | Login and Authentication issues(tech) | Security Issues (tech) | Behavioural Issues (tech) | Data issues (tech) | Refund Issues (tech) | Ads (tech) | Rewards and Offers Issues(tech) |
|---|---|---|---|---|---|---|---|---|---|
| Paytm: Secure UPI Payments | Yes | | | | Yes | Yes | Yes | | |
| Google Pay :Secure UPI payment | Yes | Yes | | | Yes | | | | Yes |
| PhonePe UPI, Payment, Recharge | Yes | Yes | | | | | | | Yes |
| BHIM MAKING INDIA CASHLESS | | Yes | Yes | | Yes | | | | |
| MobiKwik- UPI, Bills, PayLater | Yes | Yes | | | Yes | | | | Yes |
| Paytm for Business: Accept Payments for Merchants | Yes | Yes | | | Yes | Yes | | | |
| Freecharge - Pay Later, UPI | | Yes | | | Yes | Yes | | | Yes |
| PayPal - Send, Shop, Manage | | Yes | | Yes | Yes | Yes | | | Yes |
| Payzapp - UPI & Bill Payments | Yes | Yes | Yes | | Yes | | | | |

| Apps | Payment Issues (tech) | Support Issues (non-tech) | Login and Authentication issues(tech) | Security Issues (tech) | Behavioural Issues (tech) | Data issues (tech) | Refund Issues (tech) | Ads (tech) | Rewards and Offers Issues(tech) |
|---|---|---|---|---|---|---|---|---|---|
| Samsung Wallet (Samsung Pay) | | | | | Yes | | | | |
| Yuva Pay:UPI, Wallet, Payments | | Yes | | | Yes | | | | Yes |
| Money Transfer ,BHIM UPI | | | | | | | | | |
| Google Pay for Business | | | | | | Yes | | | Yes |
| iMobile Pay by ICICI Bank | Yes | | Yes | | Yes | | | | |
| PhonePe Business | | Yes | | | Yes | Yes | | | |
| Fyp- Payment App for Teens | | | | | Yes | | | Yes | |
| Skrill - Pay & Transfer Money | | Yes | Yes | Yes | Yes | | | | |
| BharatPe for Merchants | Yes | | | Yes | Yes | | | | |
| BHIM SBI Pay :Retail & Business | Yes | | Yes | | Yes | Yes | | | |
| muvin: Teenagers Payment App | | | | Yes | | | | | |
| Bhim Yes PAY - UPI, Wallet, Recharge, Bharat QR | Yes | Yes | | | | | | Yes | |

| Apps | Payment Issues (tech) | Support Issues (non-tech) | Login and Authentication issues(tech) | Security Issues (tech) | Behavioural Issues (tech) | Data issues (tech) | Refund Issues (tech) | Ads (tech) | Rewards and Offers Issues(tech) |
|---|---|---|---|---|---|---|---|---|---|
| Airtel Thanks – Recharge & UPI | Yes | Yes | | | Yes | Yes | | | |
| F-Pay, Wallet to Bank Transfer | Yes | Yes | | Yes | Yes | Yes | | Yes | |
| Razorpay Payments for Business | | Yes | | Yes | | | | | |
| Ola Money - Wallet payments | Yes | Yes | Yes | | Yes | | Yes | | |
| Bill Payment &Recharge, Wallet | Yes | Yes | | | Yes | | Yes | Yes | |
| CRED: Credit Card Bills & More | | | | | Yes | | | | Yes |
| Stripe Dashboard | Yes | Yes | Yes | | Yes | Yes | Yes | | |
| Fave - Cashback, Deals, UPI | | | | | Yes | | | | Yes |
| BHIM Axis Pay:UPI,Online Recharge & Money Transfer | Yes | Yes | Yes | | Yes | Yes | | | |

## CONCLUSION

From the above table, we can say that the most common issues is Behavioural issues with a total of 23 apps often troubling the users.Then Payment issues which hold a important role in the whole finance industry. Then data updating and being able to check data issues are common.Users are facing several issues while claiming the rewards or offers in several apps and some users are genuinely disappointed

with the rewards given in some apps. Login and Authentication too is playing a huge role in making users frustrated with so many users facing these issues in several apps.Financial Security is sometimes being at risk with some apps giving improper access controls.The at last some apps contain ads which often results in failed transactions.Refund is sometimes a lengthy and frustrating process in some apps.

| Issue | CWE error | CWE Code | Number of Apps |
|---|---|---|---|
| Payment Issues (tech) | Business Logic Errors | 840 | 16 |
| Login and Authentication issues(tech) | Authentication Errors | 1211 | 8 |
| Security Issues (tech) | Cleartext Storage of Sensitive Information | 312 | 6 |
| Behavioural Issues (tech) | Behavioral Problems | 438 | 23 |
| Data issues (tech) | Insufficient Technical Documentation | 1059 | 11 |
| Refund Issues (tech) | Business Logic Errors | 840 | 4 |
| Ads (tech) | Improper Handling of Windows ::DATA Alternate Data Stream | 69 | 4 |
| Rewards and Offers Issues(tech) | Business Logic Errors | 840 | 9 |

The colab files containing all the data scraping and also topic modeling done are below.

***Scrapper***

***Exploratory Data Analysis***

***Topic Modeling***

## REFERENCES


1)https://play.google.com/store/search?q=payment+app&c=apps

2)https://www.kaggle.com/code/prakharprasad/mobile-reviews-topic-modeling

3)https://cwe.mitre.org/