Analysis of Google Playstore Apps

Ginu Varghese
Department of Computing Science and Mathematics
Dundalk Institute of Technology
Dundalk, Co.Louth, Ireland
d00251842@student.dkit.ie

Ravichandra Reddy Kovvuri
Department of Computing Science and Mathematics
Dundalk Institute of Technology
Dundalk, Co.Louth, Ireland
d00251806@student.dkit.ie

Sujil Kumar KM
Department of Computing Science and Mathematics
Dundalk Institute of Technology
Dundalk, Co.Louth, Ireland
d00242726@student.dkit.ie

***Abstract*—Almost half of the people around the globe are using a smartphone and different applications in that phone. This report gives the analysis of different apps available in the Google play store which is the official Appstore for android operating system. This analysis focus on the cleaning of the dataset, its analysis and visualization of the apps using three different technologies the excel, power bi and tableau.**

Keywords—cleaning, excel, power bi, tableau, play store

I. Project Summary

Android is an operating system created by Google for smartphones and tablets.It is available on devices by different manufactures with different choices (J R, 2022). It has captured around 74% of the total market which is a true indicator of the huge amount of population using android. This analysis is to help android developers to know what the motivating factor for people is to download an app. It will also help to find out the factors that affect someone's decision to download an app. Here we analysed the category, reviews, price, ratings, android version and installs for this purpose and found out how they are interrelated (Sharma, 2019).

The aim of this analysis is to provide information about android applications their categories, ratings, and other related information to android users. We also analysed the data for the factors that influence an application, to know why and how certain applications succeed and others. Also, what is required for an application to be considered as successfully topping the charts. So, we used dashboards to visualize the information that can be understandable to people even with less technical knowledge. In this study we analysed data using python programming language and we use three different tools for visualisation.

The following are the research questions that were analysed through the visualization:

- Which are the popular apps in the play store?
- Do people prefer free apps or paid apps?
- Which category has the most apps, and which are the most rated apps?
- Which are the highest paid apps?

The table below shows the technology used for the project.

Technologies used
Programming language: Python (Jupyter Notebook, Visual studio)
Python libraries: • Pandas • NumPy
Visualization tools: • Microsoft Excel • Power Bi • Tableau

II. Understanding The Data

A. *Data Description*

The dataset contains the details of the different apps available in the Google play store. The dataset was downloaded from Kaggle.com. There were 13 variables and 10841 observations in the actual dataset. The dataset contains the application names its installs, ratings, and other information useful for the analysis. The android version of the apps, the late date in which the app has been updated, the price of the app and its category are also available in the dataset. We cleaned the data set using python and The final dataset contains 13 variables and 9660 observations.

Link to the dataset: https://www.kaggle.com/datasets/lava18/google-play-store-apps?select=googleplaystore.csv

B. Types of Variables

The Table II. describes the variables in the dataset. There are 13 variables in the dataset. No additional variables were created for analysis. This is the final dataset after cleaning the data:

Types of variables			
Variable Name	**Category**	**Type**	**Description**
app	Nominal Categorical	String	The application name
category	Nominal Categorical	String	The category in which the application belongs.
genre	Nominal Categorical	String	The genre of the application.
content_rating	Nominal Categorical	String	Content category of the application.
rating	Continuous Numerical	Float	Rating of the app.
reviews	Discrete Numerical	Integer	Reviews available to the application.
installs	Discrete Numerical	Integer	Number of installs or downloads done for the app.
size_in_MB	Continuous Numerical	Float	It is the size of the application in MB.
price_in_$	Continuous Numerical	Float	It is the price of the application in dollars.
type	Nominal Categorical	String	Type of the application, free or paid.
android_version	Nominal Categorical	String	Android version of the application when its released.
current_version	Nominal Categorical	String	The current version of the application.
last_updated	Date	Date	The year of the updation.

TABLE II. Type of variables in the dataset

C. Data Cleaning

The data was cleaned using python. Pandas library was used for this purpose. First, we loaded the data to the visual studio using the pandas library and removed the duplicates. Then the column names were not correct so, we changed the column names appropriately. The next task was to clean each column and set the correct data type on each column. Major cleaning is on size and genre column. The size column consists of mixed data, so we converted all the size data to kilobytes and later we converted it to megabytes. Coming to the genre column we picked the primary genre. The data consists of applications with two or more rows with the same application name we took distinct application names from the pandas data frame. After cleaning we got 9660 records out of 10841 records.

III. UNDERSTANDING THE TECHNOLOGIES

A. Microsoft Excel

Excel is very powerful tool which can create beautiful and useful dashboards. Even though the process of making graphs is quite difficult in excel compared to power BI and Tableau. There are ways we can achieve almost similar results in excel. The best part of excel is that often people are familiar with using MS office applications, so learning about excel will not be much harder for new person. Excel has limited option to produce a convenient full screen view like other technologies.

B. Power BI

Power BI is a tool designed by Microsoft to visualize the data with reasonable graphs and figures. It supports visualizations using python code also. The 3 major components in power BI are Report, Data, and Model. Using report, we can drag and drop the data needed and we can select any visualization of our preference. Data tool helps us to clean the data and we can modify the dataset. Model is used to change relationships between the data, and we can edit the relationship between tables so that we can achieve correct relationship while visualizing the data. After completing the reports, we can select graphs in the report and pin required graphs to the dashboard. We can produce dashboards for mobiles also. Power BI can be used in three different ways: power BI desktop, Power BI online, and Power BI mobile. We need some knowledge and practice to use power BI.

C. *Tableau*

Tableau is the tool for analyzing and reporting large amount of data. In tableau we can create charts and graphs in different worksheets, and we can add those to dashboard. Tableau helps in using the data easily and we can make visualizations using the variety of color palettes available in the tableau in a more interactive and efficient way. Tableau has tools that helps to data discovery and exploration which helps the users to answer important questions within no time. Tableau doesn't need any prior programming knowledge to work. users without any experience can work with tableau easily for creating visualizations. Users can incorporate multiple scripting languages to Tableau. Tableau also support responsive mobile support dashboards. Tableau doesn't provide the automatic refreshing feature to refresh the data we need to do manually if there is changes in the data. Tableau don't have option to modify data after loading so we must upload data which is cleaned properly. Tableau lacks the data modeling and data dictionary capabilities for Data Analysts. So, the user must separately maintain the metrics definitions elsewhere (Khai, 2021).

IV. Review of Dashboard

A. *Microsoft Excel Dashboard*

Strengths:

- Logo and buttons are implemented correctly.
- Switching between different dashboards to show more visualizations.
- Select all and clear filters are given for all the slicers.

Future recommendation:

- Tree map colors are odd with the dashboard.

B. *Power Bi Dashboard*

Strengths:

- Main relevant information stands out on home
- All graphs are interacting each other
- Single clear filter option is very useful Disadvantage

Future Recommendation:

- We do not normally recommend using Pie Chart. Its better to use some other graphs.

C. *Tableau Dashboard*

Strengths:

- Attractive & interactive header bar.
- It's a useful idea to have a "bulb" icon providing a short info about the dashboard.
- A color tone consistently used throughout the dashboard graphs Disadvantage

Future Recommendation:

- Genre based on installs need to be scrolled to see whole heatmap so its better to improve it.

V. Conclusion

We analyzed google playstore apps using three technologies. From our analysis we found which are the most popular apps, its ratings, and reviews. We understood that every technology has its own benefits and disadvantages. Based on the dashboard we can recommend tableau for visualizations using more data. Although it has some disadvantages it is used in most organizations and preferred by most people for visualizations since it has more options for visualization. It is a tool which we can depend for producing interactive and efficient dashboards.