**Google Play Store Analysis**

**Sumit ojha , Shafaq Khan**

**Ayush Mishra**

**Data science trainees,**

**AlmaBetter,**

**Abstract:**

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store. This journal talks about different classiﬁer mod-

els that we used for prediction purposes and ﬁnding which

one gives the highest accuracy. This journal also gives de-

tailed information on feature extraction and the complete

Data visualization done on this data set. Our project code

can be found at https://github.com/Rimshamaredia/CSCE-

421-Project.

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store. This journal talks about different classiﬁer mod-

els that we used for prediction purposes and ﬁnding which

one gives the highest accuracy. This journal also gives de-

tailed information on feature extraction and the complete

Data visualization done on this data set. Our project code

can be found at https://github.com/Rimshamaredia/CSCE-

421-Project.

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store. This journal talks about different classiﬁer mod-

els that we used for prediction purposes and ﬁnding which

one gives the highest accuracy. This journal also gives de-

tailed information on feature extraction and the complete

Data visualization done on this data set. Our project code

can be found at https://github.com/Rimshamaredia/CSCE-

421-Project.

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store. This journal talks about different classiﬁer mod-

els that we used for prediction purposes and ﬁnding which

one gives the highest accuracy. This journal also gives de-

tailed information on feature extraction and the complete

Data visualization done on this data set. Our project code

can be found at https://github.com/Rimshamaredia/CSCE-

421-Project.

Analysis of Google Play Store Data set and predict the popularity of an app on

Google Play Store

Rimsha Maredia

Texas A&M University

College Stattion, Texas

rimsha.maredia@tamu.edu

Abstract

The google play store is one of the largest and most pop-

ular Android app stores. It has an enormous amount of data

that can be used to make an optimal model. We have used a

raw data set of Google Play Store from the Kaggle website.

This data set contains 13 different features that can be used

for predicting whether an app will be successful or not using

different features. This data set is scraped from the Google

Play Store. This journal talks about different classiﬁer mod-

els that we used for prediction purposes and ﬁnding which

one gives the highest accuracy. This journal also gives de-

tailed information on feature extraction and the complete

Data visualization done on this data set. Our project code

can be found at https://github.com/Rimshamaredia/CSCE-

421-Project.

A few thousands of new applications are regularly uploaded on Google play store. A huge number of designers working freely on designing the apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application

**Introduction:**

The Play Store Apps has enormous potential to drive app-making business to success. The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product. It has used the data associated to Play Store apps for analyzing the Android market. This dataset contains details of different applications and reviews from different users,

The Google Play Store is the largest and most popular Android app store.The purpose of our project was to gather and analyze detailed information on apps in the Google Play Store in order to provide insights on app features and the current state of the Android app market.

To launch themselves successfully and create an identity for themselves in this oversaturated market, they need to ensure that majority of the essential factors are incorporated while designing and developing an app that would play an important role in customer's decision-making process. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users.

The main objective of this exploratory data analysis project is to understand customer demands better and thus help developers to popularize their product on the Play Store

## Project Files Description

This Project includes 1 colab notebook, 1 technical documentation as well as 1 presentation:

**Executable Files:**

* [Play Store App Review Analysis](https://github.com/San13deep/Play-Store-App-Review-Analysis/blob/main/Play_Store_App_Review_Analysis_Capstone_Project.ipynb) - Includes all functions required for clustering operations.
* [Google Colab](https://github.com/San13deep/Play-Store-App-Review-Analysis/blob/main/Play_Store_App_Review_Analysis_Capstone_Project.ipynb) - All the outputs are visible in the provided colab notebook.

**Input Files:**

 **Play Store Data.csv** - It contains the basic details of the app like number of user reviews, ratings, etc.

 **User Reviews.csv** - It contains the user reviews and its sentiment score for the respective app.

**Data Source:**

* [Dataset](https://learn.almabetter.com/courses/take/team-capstone-projects/texts/19443175-play-store-app-review-analysis-dataset) - Dataset taken from Almabetter

### The contents of Play Store Data are:

**App** - It tells us about the name of the application.

**Category** - It tells us about the category to which an application belongs.

**Rating** - It tells us about the ratings given by the users for a specific application.

**Reviews** - It tells us about the total number of users who have given a review for the application.

**Size** - It tells us about the size being occupied the application on the mobile phone.

**Installs** - It tells us about the total number of installs/downloads for an application.

**Type** - It tells us whether the application is free or a paid one.

**Price** - It tells us about the price of the application.

**Content Rating** - It tells us about the target audience for the application

**Last Updated** - It tells us about the when the application was updated.

**Current Ver** - It tells us about the current version of the application.

**Android Ver** - It tells us about the android version which can support the application on its platform

### The contents of User Reviews are:

* App: It contains the name of the app with a short description (optional).
* Translated\_Review: It contains the English translation of the review dropped by the user of the app.
* Sentiment: It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
* Sentiment\_Polarity: It gives the polarity of the review. Its range is [-1,1], where 1 means ‘Positive statement’ and -1 means a ‘Negative statement’.
* Sentiment\_Subjectivity: This value gives how close a reviewer’s opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer’s opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information

## Problem Statement:

1. What are the category rating above 4.5?
2. Find out the Category Wise Rating?
3. What are the top 10 categories on Play Store and No.of Apps in each category?
4. What is the percentage of free and paid apps?
5. Find out the No. of apps in each category?
6. Find out the No. of Apps in each age group?
7. What is the percentage of review sentiments

## ****What is Exploratory Data Analysis?****

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets for patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset and summarize their main characteristics, often employing data visualization methods. It is an important step in any Data Analysis or Data Science project. It helps determine how best to manipulate data sources to get the answers you need.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better and make it more attractive and appealing.

The following are the various steps involved in the EDA process:

**Problem Statement** - We shall brainstorm and understand the given data set. We shall study the attributes present in it and try to do a philosophical analysis about their meaning and importance for this problem.

**Hypothesis** - Upon studying the attributes present in the data base, we shall develop some basic hypothesis on which we can work and play with the data to look for the varied results which we can get out of it.

**Univariate Analysis** - It is the simplest form of analyzing the data. In this we would initially pick up a single attribute and study it in and out. It doesn't deal with any sort of co-relation and it's major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.

**Bivariate Analysis** - This analysis is related to cause and the relationship between the two attributes. We will try to understand the dependency of attributes on each other.

**Multivariate Analysis** - This is done when more than two variables have to be analyzed simultaneously.

**Data Cleaning** - We shall clean the dataset and handle the missing data, outliers and categorical variables.

**Testing Hypothesis** - We shall check if our data meets the assumptions required by most of the multivariate techniques.

## Steps Involved:

After loading the dataset, we can start the exploration but before that, we need to check and see that the dataset is ready for performing several exploration operations or not, so let’s first have a look at the structure and the manner in which the data is organized.

### Data Cleaning

Our data set contains a large number of null values in the rating column, so we drop them. Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column. Our data set also contain the duplicate rows for a single application. We also drop the duplicate rows because the rows contain the identical data. Also drop the rows, which have rating greater than 5.

### Data Transforming

From the information of data frame, we can see that all the columns except rating have the object data type but some of the columns like, reviews, size, installs and price have the numerical value. So, we have to transform them in proper data type and also remove the unwanted values from the numerical columns like ‘+’ and ‘,’ from installs and ‘$’ from price. In the size column we have some values in KB and some values in MB, so we transform all the values in MB.

Category

We breakdown the apps by category and observe that family and game categories have the maximum number of apps in the play store. Weather, house and home, comics, events, beauty, and parenting are the categories which have a few numbers of apps.

### Installs

We analysis the install column to observe the effect of size, price, rating, content rating, android version on app installation number. We can analysis that for each and every category number of app installation does not depend on the size. The free apps installed mostly. The apps which can be used by everyone is more installed than the apps which can be used by a particular age group. Rating of mostly installed apps is between 4 and 5.

### Exploratory Data Analysis

After establishing a good sense of each feature, we proceeded with plotting a pairwise plot between all the quantitative variables to look for any evident patterns or relationships between the features. There is a high variance in the number of installs and in number of reviews. To overcome this problem, we add two new columns to the data frame named: log\_installs and log\_review, which contain the logarithmic values of installs and review columns, respectively.

## Challenges Faced

* Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.
* We can explore the correlation between the size of the app and the version of Android on the number of installs
* The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis, if we had atleast 70% - 80% of the data available in the merged dataframes
* The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis, if we had atleast 70% - 80% of the data available in the merged dataframes

**Summary:**

Google play store Analysis project was done by group of 3 members – Shafaq Khan, Sumit Ojha and Ayush Mishra. In this project we had got 2 different csv files as an input. The csv files were play store data and user reviews. The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product. It is of 10k Play Store apps for analyzing the Android market. This dataset contains details of different applications and reviews from different users. Application distribution platform, for example, Play Store gets overwhelmed with millions of new applications being launched on the platform regularly. Lots of designers and developers work on it to make an app successful on the Play Store. In this highly competitive world, it is an immense challenge for a developer to know whether they are focusing on the right path to make their app successful on the platform.

To launch themselves successfully and create an identity for themselves in this oversaturated market, they need to ensure that majority of the essential factors are incorporated while designing and developing an app that would play an important role in customer's decision-making process. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users.

The main objective of this exploratory data analysis project is to understand customer demands better and thus help developers to popularize their product on the Play Store. In this project we analyzed the data of Google play store apps like how many number of apps are available in dataset, what is there status, there ratings, genres, reviews percentage of free and paid apps, etc. For that we used data visualization tools such as pandas ,numpy , seaborn , matplotlib. And with the help of seaborn and matpolitb we plotted the graphs and shown the distribution of apps in different age group ,percentage of free and paid apps,reviews in the apps ratings etc. The apps within the Google App store could also be biased and overrated because higher ratings given by users potentially attract several new users disproportionately. This study therefore investigated the utilization of ensemble classifiers to predict numeric ratings for Google Play store apps supported the user reviews for those apps. Several ensemble classifiers were investigated to gauge their performance on the reviews scraped from the Google App store.

**Conclusion:**

After undergoing these algorithms and process we concluded that our hypothesis is true. Meaning you can predict app ratings, However significance preprocessing must be done before you start the classification and regression processes.

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market! This shows that given the Size, Type, Price, Content Rating, and Genre of an app, we can predict about 92% accuracy if an app will have more than 100,000 installs and be a hit on the Google Play Store.

User reviews are limited to identifying polarity and subjectivity. However, the massive increase in review based data implies a requirement to focus also on performing predictions. This process is challenging yet fruitful, as user reviews are qualitative while ratings are essentially quantitative.

The numeric scoring of apps within the Google App store could also be biased and overrated because higher ratings given by users potentially attract several new users disproportionately. This study therefore investigated the utilization of ensemble classifiers to predict numeric ratings for Google Play store apps supported the user reviews for those apps. Several ensemble classifiers were investigated to gauge their performance on the reviews scraped from the Google App store.