**PLAY STORE APP REVIEW ANALYSIS**

**Avinash Yadav, Deepika Yadav**

**Data Science Trainees**

**AlmaBetter, Bangalore**

**Abstract**

Google Play is a digital distribution service operated and developed by Google Inc. It serves as the official app store for the Android operating system, allowing users to browse and download applications developed with the Android SDK and published through Google. Google Play also serves as a digital media store, offering music, books, movies, and television programs.

**Problem Statement**

Android is expanding as an operating system. It has captured around 74% of the total market which is a true indicator of the huge amount of population using android. Our goal is to help android developers to know what is the motivating factor for people to download an app? It will also help to find out the factors that affect someone’s decision to download an app. I would like to analyse category, reviews, price, ratings and installs for this purpose and find out how they are inter related.

**Goal**

The aim of our analysis is to provide insights about android applications and their categories. We want to deep dive in data for the factors of influences on 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. Our project goal is to predict the number of installs of apps by looking at app info and its reviews. We hope that this project will help app developers predict their number of installs or investors who want to pick out the next big app. Companies may run beta focus groups, or app developers may receive feedback from testers and get certain amounts of reviews. We use this and some knowledge about the app to predict its success. Knowing the number of installs can help developers and business managers because they can predict the profit. This project's result may show the importance of reviews to apps in the market as it could be one of the determining factors for the number of installs.

**Introduction:**

The dataset contains details of Android applications present on Google Play Store. For analysis of the mentioned data, three we have used Python. Our business case is to locate the best Apps, which we measure by Review check. There are 13 includes that depict each application and an aggregate of 10842 applications. Following variables were initially included:

|  |  |
| --- | --- |
| **Table Header** | **Second Header** |
| App | Name of the App |
| Category | Category of the app |
| Rating | Over all user rating of the app out of 5 on the Play Store |
| Reviews | Number of user reviews for the app |
| Size | Size of app |
| Installs | Number of user downloads/installs for the app |
| Type | Paid or Free |
| Price | Cost of the App |
| Content Rating | Age group the app is targeted at |
| Genres | An app can belong to multiple genres (apart from its main category) |
| Last Updated | Date when the app was last updated on Play Store |
| Current Ver. | Current version of the app available on Play Store |
| Android Ver. | Minimum required Android Version |

**Steps Involved**:

* **Raw Tables**:

It has information such as app name, category, rating, and more. And the other is a list of reviews for each app with the sentiment if that particular content of the review was positive, neutral, or negative. Unfortunately, we could not directly use these two files as they are not joined.

* **Data Preparation and Cleaning**:

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and the combining of data sets to enrich data. Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

* **Null Value Treatments:**

Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

* **Removing Duplicate Entries:**

We don’t want to count certain apps more than once when we analyze data, so we need to remove the duplicate entries and keep only one entry per app.

* **Standardization of features:**

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way.The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

* **Exploratory Analysis and Data Visualization**:

In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. This communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will be represented visually, determining how and to what extent the property of a graphic mark, such as size or color, will change to reflect changes in the value of a datum.

**Conclusion:**

The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project. Many other interesting possibilities can be explored using this dataset. The dataset delivers insights to understand customer demands better and thus help developers to popularize the product. Dataset can also be used to look whether the original ratings of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

**References**:

* Towards Data Science
* Blog.Jovian.ai
* GeeksforGeeks
* Kaggle.com