PLAY STORE APP REVIEW ANALYSIS

TECHNICAL DOCUMENT

Shubham Sartape

Abstract:

Play Store is one of the leading Digital Service Providers where App Developers can launch their applications. The App Developers leverage the platform by providing values to users through various tools, applications, and programs to address their multiple needs.

However, do you think it is enough for the Developers to publish their Apps and wait for the fortune/success it deserves?

Of course not, a huge analysis of these various Apps is implemented and monitored in the backend for valuable insights which are necessary for the overall success and performance of the App.

In this Project we are provided with two Data Sets. First is the 'Play Store Data', which consists of information from various Apps. Second is the 'User Reviews' Data Set which consists of User sentiment information related to the App. Leveraging these two Data Sets we have performed detailed Data Analysis and answered some of the questions which will surely help the Developers with strategies and Decision Making necessary for the growth and success of their App.

Introduction:

The Play Store is flooded with thousands of new applications regularly as the developers and designers work in and out to finally implement these applications on the platform. As we all know, most of the applications are usually Free, the monetary model is obscure or inaccessible depending upon the In-App Purchases, Memberships, user likability, installations, and reviews. Quantity of Application Installation can be a useful parameter in determining the prosperity of the App. Also, ratings and reviews are the candid feedback provided by the users and can be a useful evaluation criterion conveying the User sentiments regarding the App. However, in some cases these ratings can be biased due to insufficient/missing votes.

This study aims to provide valuable insights of various apps across Play Store. The insights are backed with detailed Data Analysis using Python Programming Language. The study consists of vibrant visualizations and accurate solutions which will help Developers with useful information needed for their Apps success.

Overview of the data:

The data we have is in a tabular format which is designed in the form of rows and columns. Rows are for the different apps and columns demonstrate different features of each app. In total there are 10841 rows and 13 columns where the first column contains the name of the apps and the other columns consist of their features. The titles of the respective columns (features of each app) are as follows:

- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content Rating
- Genres
- Last Updated
- Current Ver
- Android Ver

Analysing these columns (features) and asking the right questions (KPIs) will help us to extract amazing information.

TECHNICAL JOURNEY

From here on we will start discussing the technical aspects of our project. We will be using the Python programming language here. This is one of the most demanding languages of our generation. Many modules have been developed in Python which helps us to do several tasks in different domains.

Set the stage up:

We have got data and to use it fruitfully, for that we need to have some tools. Those tools are required to be imported first. So, we will be importing some Python modules which are as follows:

- Pandas
- Numpy
- Plotly.express
- Plotly.graph_objects
- Pyplot.subplots
- Matplotlib.pyplot
- Seaborn
- Ast

These above Python modules will be used in our work to do several things.

Next and an important part of our project are to import our Google Drive and mount it, which consequently creates a connection of our notebook to our drive. The data which we will use is saved up there in our drive. Now we can fetch the data down to our notebook from the drive easily. The

first module we will be using is pandas. Through pandas we will read the data from our drive and put it in a variable named 'data' so that we can use it down the line whenever it's needed.

Play Store data wrangling:

First, we need to know the shape of the data to get its outer structure like columns and rows. Then by using the info method we would like to get an overview of our data like the columns in it, non-null values in each of it, their data types, and its memory etc.
Then we will rename a few columns of the data as per our requirement, especially those columns which are named with two. They are renamed by just putting an underscore between them.

Fixing 'Rating'-

Now we would like to get a box plot of the data and get some insight. Surprisingly, we can see an outlier in our data where there is an app whose rating lies far away from legitimate margin '5'. looking at it we find an app with rating more than 5, which is of course kind of impossible. So that row in our data is of no use and it needs to be dropped. Now we have data where there is no app having a rating more than '5'. On looking at a box plot or histogram it seems everything is fine in terms of rating.

Cleaning 'Install'-

First, we will see the number of installs and then we see that few signs like '+' and ',' which we may find difficult in doing further calculations. So they surely are needed to be removed. So we did that and converted all those numbers into integer data type.

Filling NA Value in 'Ratings' with mode value-

There are few data in Ratings column which doesn't have any value (null value or 'NaN'). They are of no use. In fact they also create hindrance while plotting graphs. So we will fill the null value with mode value and 0. They are few apps which haven't been even installed so they also have null rating which are also useless. We filled it with mode value and 0.

Checking overall NA values in data-

It would be quite easier for us if we get to check that how many total null values is there each column. By getting the sum of all the null values, we will get the total null values of each column.

Fixing NA value in 'Type', 'Current_Ver', 'Android_Ver' with mode 0 -

In the Type, Current_Ver, Android_Ver columns, there are again few null values which are again of no use. They need to be also filled with mode 0 value again like we did before. Now on summing all the null values of each column we will get number of null values in each column.

Cleaning 'Price'-

The values in Price column are signified in unit '$' as that's the currency. But while doing calculations later it may cause trouble. So, we will remove it too and convert them into float using numpy.

Cleaning 'Size'-

The size of the apps as mentioned in the data is in different forms. There are few apps where its size is signified in 'M' which will be replaced by e+6, apps with size mentioned in 'K' will be replaced by e+3, apps with size shown as '1,000+' are being replaced by 1000. There are even few apps where no specific size is mentioned, in fact it is said as 'Varies with device'. We will replace it with null value and convert the entire data type into float using numpy.

Dropping rows with duplicate APP data –

There are many apps which are being repeated several times which may raise problem while dealing with them. So, we need to drop the duplicate apps for our ease. We will keep the first one and drop all others with the same name. This will reduce our data and make it much lighter and easier to handle.

Fixing 'Reviews' data type –

In the data we have a column named 'reviews' where the number of reviews about each app has been received is mentioned. They are of course countable numbers so they need to be in whole numbers. Hence, we have converted them into integer format.

Fixing 'Genres' by splitting at ';' and converting into list –

If we go through the 'Genres' column, we would some genres are separated by ';' which may raise concerns later while working with them. So, for our convenience we have split them at ';' and converted them into lists.

Data Exploration:

Now that we have cleaned our data sufficiently, on asking appropriate questions we can get magnificent answers which subsequently provides us useful and imperative information about the play store data.

Top Category in Play Store –

The first question we asked to the Play Store Dataset is, about the top categories in terms of number of apps. For that we will first get the size of each category which in return will give us the number of apps in specific category. On getting that we will sort their values using pandas in descending order. This will return us top categories in terms of number of apps from top to bottom.

Top 3 Categories in Play Store by Content Rating –

Here we are going to find out about the top 3 categories of apps in terms of specific content rating. We will count the values each category in every kind of content rating and will sort the values using pandas in descending order. Now by taking the first three categories in a specific content rating we will get the top 3 categories in each kind of content rating. Demonstrating it in sunburst and tree map, will help us to get a better understanding.

Top 10 most installed paid app by installation numbers and their ratings –

Most of apps in our data set are free but, few of them are few apps which are paid. We will take the app, rating and installation numbers of the apps which their type is mentioned as paid. Now by sorting the values using pandas in descending order and taking the first ten rows, we get the top 10 most installed apps by installation numbers and their ratings. Putting them in a horizontal bar graph will give a nice overview of it.

Paid v/s free apps in Play Store –

On looking at the type column we can that there are two types of apps, paid and free. On counting their values using pandas we get the number of installed apps which are paid and free respectively. Presenting it in the form of a pie chart will help us to get an overview of the distribution of paid and free apps.

Paid and free apps by category in Play Store –

By taking the apps where type is free and counting their values with respect to their categories, we get the number of free apps in each category. Similarly by taking apps where type is paid and

counting their values with respect to their categories we get the number of paid apps in each category. Getting a pie chart for both will help us a lot in understanding distribution in a great way.

App installed in each category –

We will here try to get the number of apps installed in each category. For that we will sum the installs of every category using pandas and setting up an index named 'Total Installs'. Now we a will get a data frame showing total number of apps installed in each category. Putting them into a vertical bar graph we give a nice overview of it.

Top 10 categories with rating 4.0 and above –

Here we would like to know about the number of apps in each category with a good rating which is 4 or above. In order to get that, we will count the number of apps with rating of 4 or above in each category. Then we will sort the values using pandas in descending order and show the first 10 rows in the data frame. Here's how we get the top category of apps with rating 4 or above. A vertical bar graph along with a scatter plot will give us a decent overview of it.

Top 10 categories with highest average number of reviews –

Here we will try to get the average number of reviews for each category. On getting then mean of the number reviews in each category using pandas and converting the numbers into integer type we will almost get our required result. Then we will just need to sort their values in descending order we and taking the first 10 rows from the data frame, we will get our final result here. A line chart will give us a nice idea of the result.

Paid app in each category by price and number of installations –

In each category we will take the paid apps and their prices. Then we will take the sum of total paid apps installed in each category suing pandas. Now again sorting the values in descending order we will get our required result. A scatter plot will help us to get a nice overview of it.

App installation by size –

Now we need to get the total number of apps installed of specific sizes. For that we will sum the number of apps installed of specific sizes and by sorting the values in descending order, we get our required result. Sizes of apps in the data frame are mentioned in different forms. To take the sizes on same page, we will convert it in integer and divide by 1000000. So to get a cool overview we can show it as a scatter plot.

Top App category with most reviews –

We have got a dataset about the user reviews which we haven't used much until now. Let's now merge the Play Store data with the User Review data. This gives us a broader data frame with much more useful information from it. Here we will count the number of reviews on apps in each category from the new data frame. Then by sorting their values in descending order, we will get the top category apps wit most reviews. On creating a vertical bar graph out of it, we will a clear overview of it.

Total installed apps by top 10 Genres having rating above 4.0 –

If we look at the Genres column, we will get know that its data is really cumbersome. So to get a manageable data we will first explode the Genres column data using pandas. Now we will get a data frame with ratings of specific genres along with their installs. Now we will sum the number of installs in each genre with rating more than 4 and sorting their values in descending order. Here's how we can get the total installed apps by top 10 genres having rating above 4. Putting it into a horizontal bar graph will give a nice overview of it.

User Review Data wrangling –

Along with the Play Store data we have a User Review data as well, which we used briefly a while ago to get the top category apps with most reviews. On looking at the data we can find a lot of null values here which aren't of any use for us. So we will simply go ahead and drop all of them using pandas.

Percentage of Review Sentiment –

If we look at the user review data frame, a column named 'sentiment' can be seen. Now on counting its values we can get the number of positive, negative and neutral reviews in our data. We can put it in a pie chart to get a nice view and understanding of their distribution.

Top 10 Apps with maximum positive reviews –

We just discovered the number of positive reviews in the data frame. We can count the number of apps with most number of positive sentiments. Now on sorting their values in descending order and taking the first 10 in the list, we will get the top 10 apps with maximum positive reviews. Putting it in a vertical bar graph will provide us a decent understanding of it.

Is sentiment subjectivity proportional to sentiment polarity? –

This question can be properly shown and answered by a scatter plot. Here will take the x axis as sentiment polarity and y axis as sentiment subjectivity. We can answer the above question by properly analysing the scatter plot. By analysisng the Scatter Plot we can say that Sentiment Subjectivity is not always Proportional to Sentiment Polarity. But however, in maximum cases it shows proportional behaviour when the variance is too high or low.

Conclusion:

Finally, we have reached to the end of the Analysis. We opted a step-by-step approach starting from Data Cleaning to finally providing valuable Insights. We exploited both the Datasets to gain some useful information. Cleaning the Data was quite challenging, but the learning was worth it.
We have provided some useful insights such as Top App installations, Best Category Performance, Best apps with positive reviews, etc…
We hope these insights can surely prove useful for the Developers to drive their Apps Success.