**DATASET EXPLANATION**

In the dataset we have apps which contains all the names of the diff apps and then we have the category of the apps that they belong to and we have rating of each app with the number of reviews in each column and the size of each app. In the next column we have the installation of each app and the next column contains whether the app is free or payed along with price column. The next column have the content rating and genres and next column has the dates of when it was last updated . the next columns contains current versrions and the android versions of the particular apps.

**PYTHON IPYNB EXPLANATION(google collab)**

Now we are importing all the required modules.

As Pandas is Python's popular data analysis library, it provides several different functions to visualizing our datausing Pandas for visualization is we can serialize or create a pipeline of data analysis functions and plotting functions.

The NumPy package is the workhorse of data analysis, machine learning, and scientific computing in the python ecosystem. It has the ability to slice and dice numeric data.

Seaborn helps **to visualize the statistical relationships**, To understand how variables in a dataset are related to one another and how that relationship is dependent on other variables, we perform statistical analysis. This Statistical analysis helps to visualize the trends and identify various patterns in the dataset.

Matplotlib is a Python library that is **used for plotting graphs through other libraries such as Numpy and Pandas**. Seaborn is also a Python library that is used for plotting graphs through Pandas, Matplotlib, and Numpy.

**Reading the data**

Now we read that data and store that data in googledata.

In the next cell we are going to inspect 5 rows and display the output.

.shape can be used to represent qualitative data which makes it easy to compare data both within a specific range or against other categories.

Next we describe the data.

Box plots **help visualize the distribution of quantitative values in a field**. They are also valuable for comparisons across different categorical variables or identifying outliers, if either of those exist in a dataset.we have ratings in x axis and count in y axis.

.hist is used to plot the histogram. Here we can see the histogram.

Using the .info we use read the total 13 columns .

Next to read the number of missing values in dataframe we use isnull . here all the null values are represented as the false.

To count the number of missing values In each column we use isnull and sum.here we get the total number of missing values.

Here we will check how many ratings have more than 5 outliers. Google data rating > 5 gives all the apps which have ratings more tn 5.

The drop() function return Series with specified index labels removed. It **remove elements of a Series based on specifying the index labels**.

Now all the values from range 10470 to 10474 are printed.

Here for the above new data we plot the box plot.

And similarly we plot the histogram for the above values.

Now here we are removing the columns that are 90% empty. here we read the length of the data. dropna() drops the complete row/column even if only 1 value is missing. Now we add all the new values and find the sum.

Def median , here the missing values are replaced with the median value of the entire feature column.

Now we will sum all the new median values.

Now we find the mode of type current and android verions. The mode of a set of data values is **the value that appears most often**. It is the value at which the data is most likely to be sampled.

Now we find the missing categorical values with mode. After finding the modes of all the values we add up all the values .

Now we convert the price reviews and ratings into numerical values. A lambda function is a small anonymous function.A lambda function can take any number of arguments, but can only have one expression.Here we use lambda function and convert all the values into numerical values.

Now we are going to read the first 10 values and describe it.

here groupby **allows to split the data into separate groups to perform computations for better analysis and the** agg() is **used to pass a function or list of function to be applied on a series or even each element of series separately**. In case of list of function, multiple results are returned by agg() method. And we find the sum and mean of the respective values.

Cell no.33 --- now we will plot the graph where we have the names of the apps in x axis and the ratings on y axis