**Data Cleaning of Google play Store Data Using Python-Pandas**

**Objective**: Make the data clean for predictive model for Google Play Store.

**Problem Statement:**

Google Play Store team is about to launch a new feature wherein, certain apps that are promising, are boosted in visibility.

**Domain**: General

**Analysis to be done**: Using Data cleaning identify

* Clean Empty Cells
* Clean Wrong Format
* Clean Wrong Data

**Content:**Dataset: Google Play Store data (“googleplaystore.csv”)

**Fields in the data** –

* App: Application name
* Category: Category to which the app belongs
* Rating: Overall user rating of the app
* Reviews: Number of user reviews for the app
* Size: Size of the app
* Installs: Number of user downloads/installs for the app
* Type: Paid or Free
* Price: Price of the app
* Content Rating: Age group the app is targeted at - Children / Mature 21+ / Adult
* Genres: An app can belong to multiple genres (apart from its main category). For example, a musical family game will belong to Music, Game, Family genres.
* 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 to perform**:

1. Load the data file using pandas.
2. Data Cleaning using python-pandas
   1. **Clean Empty Cells**
3. Check for null values in the data. Get the number of null values for each column.
4. Drop records with nulls in any of the columns.
   1. **Clean Wrong Format**
5. Variables seem to have incorrect type and inconsistent formatting. You need to fix them:
6. Size column has sizes in Kb as well as Mb. To analyze, you’ll need to convert these to numeric.
   1. Extract the numeric value from the column
   2. Multiply the value by 1,000, if size is mentioned in Mb
7. Reviews are a numeric field that is loaded as a string field. Convert it to numeric (int/float).
   1. **Clean Wrong Data**
8. Variables seem to have incorrect type and inconsistent formatting. You need to fix them:
9. Installs field is currently stored as string and has values like 1,000,000+.
   1. Treat 1,000,000+ as 1,000,000
   2. remove ‘+’, ‘,’ from the field, convert it to integer
10. Price field is a string and has $ symbol. Remove ‘$’ sign, and convert it to numeric.
    1. **Sanity checks:**
11. Average rating should be between 1 and 5 as only these values are allowed on the play store. Drop the rows that have a value outside this range.
12. Reviews should not be more than installs as only those who installed can review the app. If there are any such records, drop them.
13. For free apps (type = “Free”), the price should not be >0. Drop any such rows.