**Headaches:** -

1. Data file is taking 9.5 GB of ram takes time to be operable and bogs down the system.
2. Data file was in an unsupported codec, had to save the file again in Unicode Transformation Format 8 (UTF – 8) to load into the notebook.
3. Got this error: have mixed types. “Specify dtype option on import or set low\_memory=False”. We pass “low\_memory=False”

It is guessing the dtype of each column which it does only after reading the whole file which is very memory demanding.

1. Still had issues trying to read the data file and used the following code to read the files in chunks:

mylist = []

for chunk in pd.read\_csv(‘games.csv', sep=';', chunksize=20000):

mylist.append(chunk)

big\_data = pd.concat(mylist, axis= 0)

del mylist

* The data was read only halfway and again ran out memory.
* Checked the shape, had 9000 columns out of which only 30 were actual column rest were just empty fields with empty records.
* Cleared out the blanks, which helped me load the data without any issues and the file size reduced from 950MB to 100MB.

1. Clean the data file.
   * Dropping columns that we don't need.
   * *Cleaning AppID:*   
     Looks like these are empty or intangible rows hence we will drop it (16 Records)
   * *Cleaning Name:*   
     Dropped 6 records with null values.
   * *Cleaning Estimated owners:* (1 Record)  
     Dropped 1 record of corrupted data.
   * *Cleaning Supported Languages:*Dropped games which were cancelled. (6 Records)
   * *Cleaning Categories & Genre:*  
     Some of the games were just Playtest, Third party servers or Utilities which we dropped (4,712 Records)
2. Now we load Games Details into PowerBI:
   * We convert Release date column from string to date dtype. (Custom Locale)
   * We make a custom column “Estimated Avg Owners” on the basis of “Estimated owners” using the following formula (M Language):-

let

// Split the range into two parts

rangeParts = Text.Split([Estimated owners], " - "),

// Convert each part to a number

startNumber = Number.FromText(rangeParts{0}),

endNumber = Number.FromText(rangeParts{1}),

// Calculate the average of the range

average = (startNumber + endNumber) / 2

in

average

* + We merge all the OS columns into 1 and remove the “,” error.
  + Don’t Summarize Estimated Owners and Price
  + Grouping Playstyle

1. Loading Game Length into Power:
   * We Delete Unwanted Columns.
   * Changed “sid” header to “AppID” to help make a connection later on.
   * Merged Meta Score on condition of column named “Merged Meta Score”.

if [meta\_uscore] <> null then [meta\_uscore] else [meta\_score]

* + Added a column named “gfq (s100)” which scaled the rating to out of 100.

Number.Round([gfq\_rating] \* 20)

* + Merged it with Merged Meta Score and renamed it to “Rating” and deleted other ratings columns.
  + Same thing with “store\_uscore”.
  + Some column name changes.

1. Genres Column
   * Applying Contains Character condition for filter usage.
2. Gaming Habits:
   * Grouping Playstyles