# Data Cleaning/Preprocessing of MetaData

## Removing tables

There were quite a large number of empty rows caused by HTML tables that messed up the csv file and caused many rows to be generated for just one row of actual data:

Chart

Description automatically generated with medium confidence

Rows removed & $24.90 & ASIN put back in proper cells (Q3382 and R3382)

## Removing problem entries

1. Some empty cells had ‘[ ]’ instead of being empty. All of these cells were converted to empty cells

Graphical user interface, application, table, Excel

Description automatically generated

1. Product descriptions were modified also to be easily processed in SQL & python. Starting characters like [‘ were removed

Graphical user interface, application, Word

Description automatically generated

## Ranks

Ranks were cleaned up:

1. The ranks had additional text which was removed:

“2,938,573 in Beauty & Personal Care (” 🡪 “2938573”

1. Miscategorized products/rank rows were removed (73 in total out of 32981 entries)

Text, table

Description automatically generated

## Prices

Prices were cleaned up:

1. Dollar amounts were converted to float (to make data handling easier): $44.99 🡪 44.99
2. Problem text strings in price columns were removed: “.a-box-inn” removed

Table

Description automatically generated

## Also View & Also Buy:

Lists were cleaned up. Example:

[‘B01B8BR0O8’, ‘B01B8BR0NO’, ‘B014MHXXM8’] 🡪 B01B8BR0O8, B01B8BR0NO, B014MHXXM8

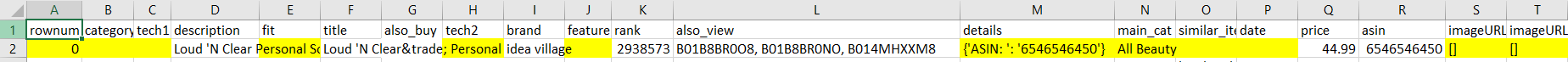
Graphical user interface, application

Description automatically generated with medium confidence

## Removal of Unused Columns

Columns removed (not needed or too sparse to be used):

* rownum
* category
* tech1
* fit
* tech2
* feature
* details
* main\_cat
* similar\_item
* date
* imageURL
* imageURLHighRes



# Data Prepping of MetaData

See [MetaDataProcessing.ipynb](file:///C:\Users\leslie_major\Documents\BAIT580A\Project\deliverable2\MetaDataProcessing.ipynb)

## ASINs & Products Table Creation

* After loading MetaData in Python – ran .info() and .nunique on dataframe
* Discovered that 352 asin’s in metadata are not unique.
* Removed those rows that had non-unique asins (kept the first row with the unique asin and deleted any subsequent rows with same asin)
* .drop\_duplicates(subset ="asin", keep = False, inplace = True) was used to do this
* 704 non-unique rows were dropped
* Remaining rows all had non-empty ranks also
* Almost all remaining rows all had unique titles also
* 31512 unique rows/products were kept

# Product Table Info:

[See products.csv](products.csv)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **# Unique** | **# Non-Null** | **# Null** | **Type** |
| rownum | 31512 | 31512 | 0 | Integer |
| asin | 31512 | 31512 | 0 | Text (primary key) |
| rank | 31512 | 31512 | 0 | Integer |
| description | 13206 | 14278 | 17234 | Text |
| title | 31331 | 31512 | 0 | Text |
| brand | 7644 | 16303 | 15209 | Text |
| price | 3461 | 10936 | 20576 | Float |

This data was now ready to be inputted into a table in SQL. The csv file ‘[**products.csv**](products.csv)’ was created in Python.

# Data Cleaning of Review Data:

## asin

‘asin’ column values all converted to text (some were coded as integer)

## overall

‘overall’ values were all converted to integer

## verified

all verified values were coded as Boolean (no change needed)

## reviewerID

all reviewerID values were coded as text (no change needed)

## reviewText

all reviewerID values were coded as text (no change needed)

## summary

all summary values were coded as text (no change needed)

## unixReviewTime

all unixReviewTime values were coded as integer (no change needed)

## vote

blank votes were coded as text instead of integer- there were changed to integer type

## style

no change needed

# Processing of Review Data

See <ReviewDataProcessing_wPolarity.ipynb>

## Review Text Column Processing

* Text strings were tokenized using word\_tokenize from nltk package
* Stopwords were removed from text
* Sentiment analysis was performed and polarity score was returned (as a new column “**Review\_polarity**”)

## Summary Text Column Processing

* Text strings were tokenized using word\_tokenize from nltk package
* Stopwords were removed from text
* Sentiment analysis was performed and polarity score was returned (as a new column “**summary\_polarity**”)

New CSV files were outputted:

* '<review_data_w_polarity.csv>' (used to make “ReviewData Table”) – see below:
* '<review_data_w_polarity_tokens.csv>'

The second file also includes the processed text data (as tokenized lists with stopwords removed) in case more analysis might be done on text data in the future.

# ReviewData Table Info:

See [review\_data\_w\_polarity.csv](file:///C:\Users\leslie_major\Documents\BAIT580A\Project\deliverable2\review_data_w_polarity.csv)'.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column** | **# Unique** | **# Non-Null** | **# Null** | **Type** |
| rownum | 371345 | 371345 | 0 | Integer (primary key) |
| asin | 32586 | 371345 | 0 | Text |
| overall | 5 | 371345 | 0 | Integer |
| verified | 2 | 371345 | 0 | Boolean |
| reviewerID | 324038 | 371345 | 0 | Text |
| reviewText | -- | 370946 | 399 | Text |
| summary | -- | 371134 | 211 | Text |
| unixReviewTime | 5288 | 371345 | 0 | Integer |
| vote | 5 | 51899 | 319446 | Float |
| style | 4780 | 125958 | 245387 | Integer |
| Review\_polarity | 41728 | 371345 | 0 | Float |
| summary\_polarity | 2501 | 371345 | 0 | Float |