IIW – 5 Data Cleaning

Dataset from the BestBuy.com

1) **Field** : URL
   Description : URL of the product.
   Operation : Unfolding operation was performed on this field to unfold/extract the unique product id of each product. A new column called product_id was created which have product id from the URL. Initially the product id was part of the url. Add a column based on current column operation was performed. Below is the regex/operation performed: value.match(/.*?(id=\w+).*?/)[0].replace("id=","")
   Edit Column-> Add Column based on current column.
   Example :
   Before :

```
{
  "URL" : "http://www.bestbuy.com/site/home-appliances/appliance-parts-accessories/abcat0916000.c?id=abcat0916000",
  "product" : "Appliance Parts & Accessories",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "N/A"
}
```

   After :

```
{
  "URL" : "http://www.bestbuy.com/site/home-appliances/appliance-parts-accessories/abcat0916000.c?id=abcat0916000",
  "product_id" : "abcat0916000",
  "product" : "Appliance Parts & Accessories",
  "model" : "N/A",
  "sku" : "N/A",
  "rating" : "0.0",
  "price" : "N/A"
}
```

2) **Field** : Price
   Description : Price of the product.
   Operation: The current records are in the form of 5, 60.5, 50.99 and empty record. Standarized the record so that it has format of two decimal values and N/A if the record is empty. The operation performed is Text Facet to convert into appropriate format.
   Facet->Text Facet
   Example:

Before:

{
"URL" : "http://www.bestbuy.com/site/buying-guides/regularcatpcmcat325300050016/pcmcat325300050016.c?id=pcmcat325300050016",
"product_id" : "pcmcat325300050016",
"product" : "Vacuum Cleaner Buying Guide",
"model" : "N/A",
"sku" : "N/A",
"rating" : "0.0",
"price" : "60"
}

After:

{
"URL" : "http://www.bestbuy.com/site/buying-guides/regularcatpcmcat325300050016/pcmcat325300050016.c?id=pcmcat325300050016",
"product_id" : "pcmcat325300050016",
"product" : "Vacuum Cleaner Buying Guide",
"model" : "N/A",
"sku" : "N/A",
"rating" : "0.0",
"price" : "60.00"
}

3) Field : Model_No

Description: The Model number of the product.

Operation: The model number had some incorrect data like “Some sentence” which was not a model number. Hence all the model number which had spaces require removed as incorrect Model no. Transform operation was performed to achieve this.

Text Filter(\.*?\s+.*)->Text Filter

Example:

Before:

{"URL": "http://www.bestbuy.com/site/home-appliances/ranges-ovens-stoves/abcat0904000.c?id=abcat0904000", "product": "Ranges, Cooktops & Ovens", "model": "Shop All \u203a"}

After:

{
"URL" : "http://www.bestbuy.com/site/home-appliances/ranges-ovens-stoves/abcat0904000.c?id=abcat0904000",
"product_id" : "abcat0904000",
"product" : "Ranges, Cooktops & Ovens",
"model" : "N/A",
"sku" : "N/A",
"rating" : "0.0",

"price" : "N/A"
}

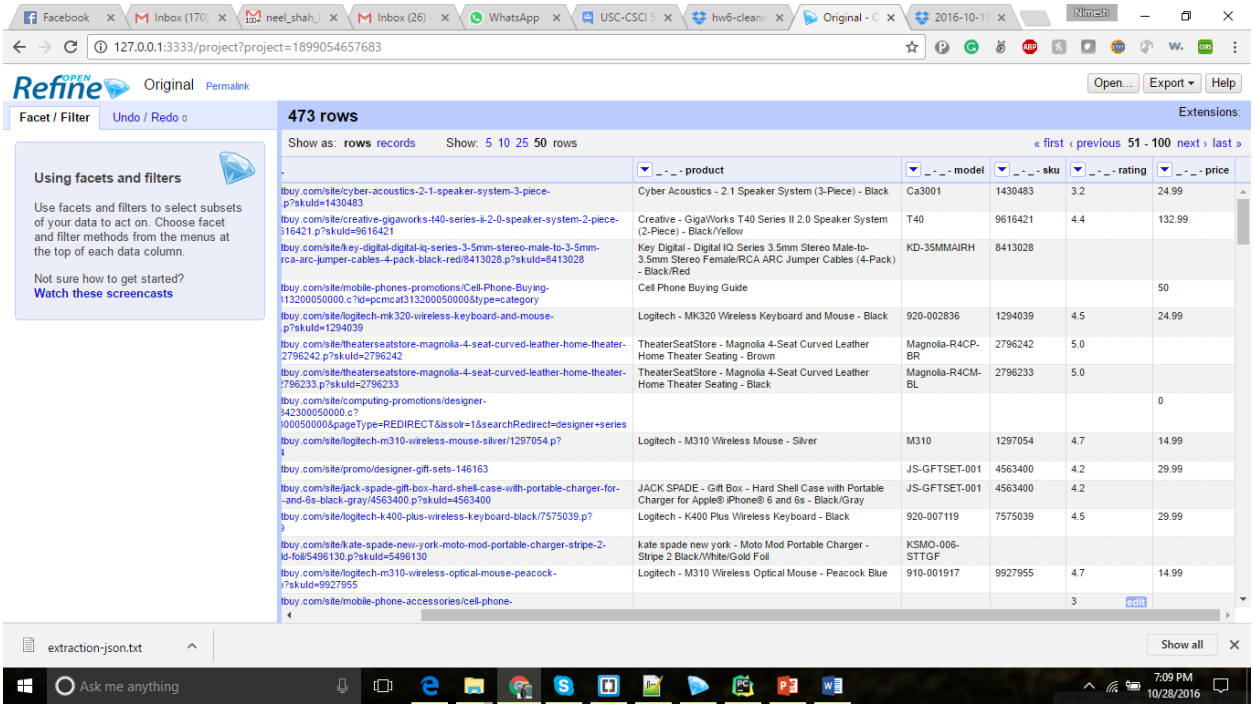
4) **Field** : SKU
Description: SKU number of the product.
Operation: Whichever blank SKU number were there they were converted to N/A using facet operation.
Facet->Text Facet
Example:
Before:
The SKU field was missing.
After:

{
"URL" : "http://www.bestbuy.com/site/home-appliances/ranges-ovens-stoves/abcat0904000.c?id=abcat0904000",
"product_id" : "abcat0904000",
"product" : "Ranges, Cooktops & Ovens",
"model" : "N/A",
"sku" : "N/A",
"rating" : "0.0",
"price" : "N/A"
}

5) **Field** : Rating
Description: Rating of the product.
Operation: Text Facet was performed for empty record as "0.0"
Facet->Text Facet.
Example:
Before:
There was no empty rating field for this record.
After:

{
"URL" : "http://www.bestbuy.com/site/buying-guides/regularcatpcmcat325300050016/pcmcat325300050016.c?id=pcmcat325300050016",
"product_id" : "pcmcat325300050016",
"product" : "Vacuum Cleaner Buying Guide",
"model" : "N/A",
"sku" : "N/A",
"rating" : "0.0",
"price" : "60.00"
}

Before:



After: