In [1]:
```
import os
import pandas as pd
pd.options.display.float_format = '{:20,.2f}'.format
pd.set_option('display.max_rows', 5000)
pd.set_option('display.max_columns', 5000)
pd.set_option('display.width', 1000)
pd.set_option('display.max_colwidth', -1)
```

In [2]:
```
walmart_amazon_path = (r'/home/ubuntu/jupyter/Hadoco/1_Standard Data In
                       r'/Unprocessed Data/product_samples/walmart_amazon')
```

Walmart Dataset

In [3]:
```
wal_fields = ['custom_id', 'title', 'brand','longdescr', 'price', 'grou
```

In [4]:
```
walmart_data = pd.read_csv(
    os.path.join(walmart_amazon_path, 'walmart.csv'),
    sep = ',',
    usecols = wal_fields,
)[wal_fields]
```

In [5]:
```
walmart_data.rename(
    columns = {
        'custom_id': 'record_id',
        'title': 'name',
        'brand': 'producer',
        'longdescr': 'description',
        'groupname': 'category'
    },
    inplace = True
)
```

In [6]:
```
walmart_data.columns
```

Out[6]:
```
Index(['record_id', 'name', 'producer', 'description', 'price', 'categ
ory'], dtype='object')
```

In [7]:
```
walmart_data.head()
```

Out[7]:

	record_id	name	producer	description	price	category
0	1	Draper Infrared Remote Transmitter	Draper	DR1143Infrared transmitter. 3-button operation for instant access to up down and stop functions. Fully compatible with learnable IR master control systems. Receiver sold seperately plugs into the Draper low-voltage control unit LVC-III sold separately .	58.45	Electronics - General
1	2	Epson 1500 Hours 200W UHE Projector Lamp ELPLP12	Epson	EPSON ELPLP12 1500HRS 200V REPL LAMP FOR LAMP POWERLITE FOR 7700P 5600P 7600 Features Lamp Life 1500 Hour Manufacturer Epson Corporation Compatible Devices LCD Manufacturer Part Number ELPLP12 Manufacturer Website Address www.epson.com Product Name Replacement Lamp Package Type Retail Product Type 200W UHE Projector Lamp Tech Specs Manufacturer Epson Corporation Manufacturer Part Number ELPLP12 Shipping Dimensions 5.25 Depth Manufacturer Website Address www.epson.com Lamp Life 1500 Hour Compatibility Epson Powerlite 7700P Projector Epson Powerlite 7600P Projector Epson Powerlite 5600P Projector Compatible Devices LCD Product Name Replacement Lamp Shipping Weight 1 lb Package Type Retail Product Type 200W UHE Projector Lamp	438.84	Monitors

	record_id	name	producer	description	price	category
2	3	Comprehensive Two-Piece 75 Precision BNC Jack for RG-59 Set of 25	Comprehensive	CH1151Comprehensive s True 75 connectors eliminate impedance mismatching and distortion that can be caused by using 50 connectors on 75 cable. True 75 BNC connectors are a must for all digital high bandwidth applications above 300 MHz and recommended for all analog applications. A lifetime warranty makes Comprehensive connectors the preferred choice of video multimedia consultants and engineers worldwide. Specifications Electrical Impedance 75 ohm Freq. Range 0-4 GHZ w low reflection 500V peak 1.3 max. Thermal limits Usable to 11 GHZ Voltage Rating PE insulators -55 degrees C to 85 degrees C Mechanical Mating 2-stud bayonet lock coupling per m39012 Center Contact Captivated center contact Cable Retention Crimps 20 to 50 lbs Material Center contacts Male Female Brass beryllium copper Body Brass astroplate finish Crimp Ferrule Copper Insulators TFE Teflon 75 all crimps rexolite Derlin 50 glass IFE hermetically sealed Clamp Gaskets Synthetic rubber silicone rubber Boots O ring crimp silicone rubber Heat Shrink Tubing All other crimps thermofit plastic Comprehensive offers a lifetime warranty on all products	59.25	TV Accessories
3	4	D-Link DCS-1100 Network Camera	D-Link	Surveillance Network Camera Built-in Omni-directional Microphone Supports 4 profiles simultaneously Adjustable image size and quality Time stamp and text overlay Three configurable motion detection windows Flip and Mirror	99.82	Garden - General
4	5	StarTech.com RKPW247015 24 Outlet Power Strip	StarTech	24 Outlet Power Strip solution for your rackmount equipment Fits most 42U racks and equipment cabinets 10 ft. power cord provides plenty of reach to power outlets 24 evenly spaced outlets help keep power cabling clean and efficiently routed	59.00	Electronics - General

In [8]:
```
walmart_data['category'].nunique()
```

Out[8]: 63

Amazon Dataset

In [11]:
```
az_fields = ['custom_id', 'title', 'brand', 'proddescrlong', 'price', '
```

In [12]:
```
amazon_data = pd.read_csv(
    os.path.join(walmart_amazon_path, 'amazon.csv'),
    sep = ',',
    #usecols = az_fields,
)#[az_fields]
```

In [15]:
```
amazon_data['category1'].value_counts()
```

Out[15]:

Computers Accessories	1147
Laminating Supplies	639
Headphones	482
Storage Presentation Materials	467
Inkjet Printer Ink	456
Bags Cases	429
Mice	402
Audio Video Accessories	394
Cases Sleeves	394
Point Shoot Digital Cameras	389
Projection Screens	370
USB Flash Drives	360
Cases	348
Cases Bags	317
Laser Printer Toner	273
Memory	261
Printer Ink Toner	247
Accessories Supplies	226
Blank Media	212
Cables Interconnects	198

In [16]:
```
amazon_data.head()
```

Out[16]:

	custom_id	url	asin	brand	modelno	categor

In [13]:
```
amazon_data.columns
```

Out[13]:
```
Index(['custom_id', 'url', 'asin', 'brand', 'modelno', 'category1', 'p
category1', 'category2', 'pcategory2', 'title', 'listprice', 'price',
'prodfeatures', 'techdetails', 'proddescrshort', 'proddescrlong', 'dim
ensions', 'imageurl', 'itemweight', 'shipweight', 'orig_prodfeatures',
'orig_techdetails'], dtype='object')
```

In [57]:
```
amazon_data.rename(
    columns = {
        'custom_id': 'record_id',
        'title': 'name',
        'brand': 'producer',
        'proddescrlong': 'description',
        'category1': 'category',
    },
    inplace = True
)
```

In [58]:
```
amazon_data.columns
```

Out[58]:
```
Index(['record_id', 'name', 'producer', 'description', 'price', 'categ
ory'], dtype='object')
```

In [59]:
```
amazon_data.head()
```

Out[59]:

	record_id	name	producer	description	price	category
0	1	Koss EQ50 3-Band Stereo Equalizer	Koss	The pocket-size Koss 3-Band Equalizer delivers high-fidelity performance and output normally reserved for more expensive home systems. With a 10dB boost or -10dB cut range of level it features a 3-band equalizer that allows for convenient and individual bass midrange and treble adjustment. Power output is greater than 20mW per channel providing clean and undistorted output into your favorite stereophones. Ergonomically designed for easy handling a rotary volume control and on off switch are placed for convenient usage.	12.65	Headphone Accessories
1	2	Kodak Black Ink Cartridge 10B 1163641	Kodak	Kodak Black Ink Cartridge 10B is a standard black ink cartridge	10.28	Inkjet Printer Ink
2	3	Kingston 128MX64 PC2700 COMPAQ Evo D320 KTC-D320 1G	Kingston	1GB - 333MHz DDR333 PC2700 - DDR SDRAM - 184-pin DIMM	33.75	Computers Accessories

	record_id	name	producer	description	price	category
3	4	Kinamax MS-UES2 Mini High Precision USB 3-Button 3D Optical Scroll Wheel Mouse with Retractable Cable Black	Kinamax	KINAMAX MS-UES2 Mini High Precision USB 3-Button 3D Optical Scroll Wheel Mouse with Retractable Cable This USB Optical Mini Mouse 800 DPI is an excellent tool for mobile professionals. The 800 dpi resolution offers twice the accuracy of standard mice and is ideal for graphics and web designers who demand pinpoint accuracy. It is small lightweight consumes virtually no room in laptop carry-bags works on most surfaces and needs just inches for accurate movement. The included retractable USB cord helps avoid extra cable cluttering your workspace by adjusting the retractable cord to just the length you require. Furthermore this compact mouse fits easily into a travel pouch for easy storage. The ergonomic contour design offers comfortable use for hours moreover its greater speed reduces excessive and potentially harmful arm movement. This Mini Mouse also features a retractable cable that is quick and easy to manage without the tangles and mess of standard cables. Its optical technology provides precise tracking on almost any surface. Just plug into the USB port and its works within seconds. This unit is USB ready Plug-n-Play and Hot Pluggable. Features Convenient Roller Button The Roller Button can be used for fast browsing up down left right on the internet in Windows documents. as well as zooming in Microsoft Office and Internet Explorer. Sleek and ergonomic designed mouse With a sleek contoured shape in classic. elegant color the PowerScroll Point is a comfortable fir for both hands and for long time operation. Plus it provides high precision mouse movements for the internet multimedia or latest Windows environments High precision optical sensor sends 1 500 signals per second to digital signal processor to determine exact location so it never needs a mouse ball like a traditional mouse. Plus it eliminates erratic cursor movement and dust accumulation.	6.99	Mice
4	5	Kensington K72349US Wireless Mouse for Netbooks	Kensington	Wireless MOUSE FOR NETBOOKS USBWRLS	24.00	Mice

Matches Walmart - Amazon

In [30]:
```
wal_az_match_data = pd.read_csv(
    os.path.join(walmart_amazon_path, 'matches_walmart_amazon.csv'),
    sep = ',',
    comment = '#'
    #usecols = ebooks1_fields,
)#[ebooks1_fields]
```

In [31]:
```
wal_az_match_data.rename(
    columns = {
        'id1': 'walmart_id',
        'id2': 'amazon_id'
    },
    inplace = True
)
```

In [32]:
```
wal_az_match_data.head()
```

Out[32]:

	walmart_id	amazon_id
0	395	1768
1	1234	12408
2	1341	8497
3	1093	13748
4	1719	12857

In [63]:
```
walmart_data[walmart_data['record_id'] == 1719]['name'].to_string(index=
```

Out[63]:
```
' Kodak ESP 3250 All-in-One Printer Value Bundle'
```

In [62]:
```
amazon_data[amazon_data['record_id'] == 12857]['name'].to_string(index=I
```

Out[62]:
```
' Kodak ESP 3250 All-in-One - Multifunction printer copier scanner - c
olor - ink-jet - copying up to 30 ppm mono  29 ppm color - printing up
to 30 ppm mono  29 ppm color - 100 sheets - Hi-Speed USB'
```

In [72]:
```
amazon_data['source'] = 'amazon'
walmart_data['source'] = 'walmart'
```

In [73]:
```
az_wal_all = pd.concat([walmart_data, amazon_data])
```

In [74]:
```
az_wal_all.rename(columns={'record_id': 'Id'}, inplace=True)
```

In [75]:
```
products_all_path = (r'/home/ubuntu/jupyter/Hadoco/1_Standard Data Inte
                     r'/Processed Data/product_samples/')
```

In [76]:
```
az_wal_all.to_csv(products_all_path + 'amazon_walmart_all.csv', sep = '
```

In [ ]:
```

```