CS784 – PROJECT STAGE-1

Data Cleaning and Understanding

Abhinav Mehra (amehra3@wisc.edu)
Danish Khan (dkhan@wisc.edu)
Mehreen Ali (mali24@wisc.edu)

List of all attributes and their frequency (Total count = 585)

Attribute Name	Frequency	Attribute Name	Frequency
Product Type	34491	California Residents Prop 65 Warning Required	806
Product Name	34491	Electronics Certifications	749
Product Segment	34491	Release Date	747
Product Long Description	34304	Material	743
Brand	27461	Digital Video Formats	679
Product Short Description	17977	Operating System Required	675
GTIN	16994	Warranty Provider	675
UPC	16684	Solid-State Drive	652
Country of Origin: Components	14886	RAM Memory Type	652
Category	13569	Wireless	590
Manufacturer Part Number	12725	Touchscreen	563
Warranty Information	12659	Video Game Platform	543
Manufacturer	8440	Monitor Type	530
Color	8017	Computer Mouse Included	527
Actual Color	8016	Keyboard Included	527
Assembled Product Length	7893	Computer Mouse Type	527
Assembled Product Width	7860	Wireless Keyboard Included	527
Assembled Product Height	7743	Desktop Computer Type	520
Warranty Length	3942	Compatible Devices	415
Condition	2773	Publisher	414
Composite Wood Code	2669	Size	364
E-Waste Recycling Compliance Required	2425	Interface Type	285
Type	2370	ISBN-13	247
CPSC-Regulated Indicator	2243	Number of Pages	245
Has Mercury	1863	Battery Life	241
Operating System	1638	Abridged	238
Multipack Indicator	1539	Bluetooth	204
Battery Type	1535	Batteries Included	194
Screen Size	1466	Resolution	187
Features	1417	Display Resolution	183
Number of Batteries	1315	Recommended Use	174
Hard Drive Capacity	1274	Genre	171
Processor Type	1235	Online Multiplayer Available	167
Energy Star	1112	Downloadable Content Available	167
RAM Memory	1010	Cooperative Gameplay Available	167
Memory Capacity	917	Subgenre	166
Processor Speed	887	Graphic Card	165
Connector Type	827	Assembly Required	165
Processor Core Type	816	Compatible Brands	164

Depth	163	Total Units	68
Display Technology	162	Audio/Video Cable Type	64
Volts	160	Frequency Range	63
Peripherals Required	155	Earpiece Type	62
Internal/External	152	Number of Channels	60
Aspect Ratio	151	Output Power	60
Average Customer Rating	145	Amps	59
Product Accessories Included	132	Processor Brand	59
Rack Size	131	Recommended Screen Size	58
Energy Guide: Appliance Labeling Rule Required	127	Compatible Models	58
Number of Audio Inputs	126	Capacity	56
Package Quantity	125	Base Unit of Measure	54
Widescreen	125	Remote Control Included	54
Is Wireless Microphone	122	Maximum Number of Online Players	53
Finish	112	Age Group	52
System Requirements	112	Record Label	52
Supported Media Formats	107	Audio Decoder	48
Lifestage	104	Enclosure Color	48
Form	104	Volume Unit	47
Created By	102	Maximum Recommended Age	45
Modified By	102	Minimum Recommended Age	45
Created Date	102	Has Parental Controls	44
Modified Date	102	Cord Length	44
Refresh Rate	102	Languages	42
Packaging Instructions	97	Write Speed	42
Dot ID	94	Number of Speakers	41
Maximum Weight	93	Battery Size	40
HDMI Connector	88	Frequency Response Range	40
Mount Type	86	Maximum RAM Supported	40
Batteries Required	79	Electronics Carrying Case Type	40
Wireless Technology	79	Number of Poles	38
Watts	79	Adjustable Tilt	38
Digital Audio Formats	78	Analog TV Tuner	38
Device Type	78	Style	38
Gender	77	Impedance	36
Maximum Number of Offline Players	75	Save Big	36
Series Title	72	Output Mode	36
Technology	70	Cable Length	34
Internet Protocol	68	Read Speed	34

Mounting Pattern	34	Department of Transportation Hazard Class	21
Response Time	33	Headphones Included	20
Shielding Material	33	Contrast Ratio Range	20
Maximum Watts Per Channel	33	Maximum Load Weight	19
Number of Outlets	32	Refillable	19
Available Instant Content Sources	32	Hard Drive Type	19
Power Type	32	Recommended Location	18
Automatic Document Feeder	32	Maximum Monthly Volume	18
Duration	32	Brightness	18
Video Game Collection	31	Shipping Weight	18
MPAA Rating	31	On-Screen Display	18
Paper Size	31	Recording Time	17
Connector Gender	31	Actors	17
Maximum Energy Surge Rating	30	Sports Team	17
Data Transfer Rate	30	DVD Region Code	17
Surge Suppression	30	Product Dimensions	17
Power Rating	30	Signal-to-Noise Ratio	17
Musician	30	Number of Megapixels	17
Instruction Manual Included	29	Has Circuit Breaker	17
Studio & Production Company	29	Shipping Height	16
Operating Power Consumption	29	Sensitivity	16
Fastener Type	29	Shipping Depth	16
Diameter	28	Clamping Level	16
In-Line Microphone	28	Humidity Range	16
Number of Ports	27	Quality Tested for Walmart	16
Age Range	26	Internal Memory	16
Automatic Two-Sided Printing	25	Shipping Width	16
DIN Size	25	Diagonal Size	16
Maximum Print Resolution	25	Number of Drive Bays	16
Minimum Operating Temperature	25	Pattern	16
Maximum Operating Temperature	25	Compatible Cameras	15
Number of Sheets	25	Made in Country	15
Items Included	24	Response Bandwidth	15
Number of Discs	24	Total Harmonic Distortion	15
Printing Speed	23	Headset Type	15
Version	22	Vehicle Model	15
TV & Monitor Mounting Components	22	Color Pages Per Minute	15
Supported Networking Standards	22	Data Line Protection	15
Instruction Manual Languages	22	Input Connector Type	14

Speaker Driver Types	14	Book Type	9
Noise Filtration	14	Has Expiration	9
Personalizable	14	Wire Gauge	9
Maximum Output	14	Top Mount Depth	9
Number of Pieces	14	Has Cooling Fan	9
Apps Installed	14	Video Streaming Quality	8
Title	13	High Pass Frequency Range	8
Talk Time	13	Gifts by Recipient	8
Enclosure Type	13	Percentage of Preconsumer Content	8
Cell Phone Service Provider	13	Digital Camera Type	8
Theme	12	Shape	8
Age Restriction	12	Surround Sound Mode	8
Thermal Management Type	12	Attachment Style	8
Director	12	Number of HDMI Connections	8
Messaging Supported	12	Number in Series	8
Distance From Wall	12	Printer Cartridge Type	8
Multifunctional	12	Percentage of Postconsumer Content	8
Low Pass Frequency Range	12	Cordless	8
Maximum Page Yield	12	Has Warranty	8
Gain Level	12	Cord Material	7
HDTV	12	Recordable Media Formats	7
Bus Speed	11	Bass Boost Frequency	7
Occasion	11	Bottom Mount Depth	7
Software Included	11	Target Audience	7
Maximum Data Transfer Rate	11	Number of Line Sources	7
Volume Capacity	11	SD Speed Class	7
USB Version	10	Television Type	7
Backlight Type	10	Battery Backup	7
Media Load Type	10	Number of Shelves	6
Frequency	10	Compatible Cars	6
Speed	10	Computer Software Type	6
Battery Watt Hour	10	Portable	6
Number of Presets	10	Remote Control Type	6
Battery Weight	10	Stereo Reception System	6
Viewing Angle	10	HDCP Compatible	6
Microphone Included	10	Equalizer Type	6
ESRB Rating	9	Upscaling	6
Video Modes	9	Has Headphone Jack	6
Hardware Included	9	USB Port	6
Cold Crank Amp	9	Crossover Slope	6
Character	9	Music Media Format	6
Number of Equalizers	9	Flash Modes	6

Total Pixels	5	Has Face Detection	3
Image Sensor	5	Portable Radio Type	3
Maximum Video Bandwidth	5	Image Stabilization Type	3
Focal Length	5	Remote Controlled	3
Has Phase Shift Selector	5	Sensor Resolution	3
Adjustable Height	5	Magnification	3
Input Signal Voltage	5	Primary Distributor ID	3
Diaphragm Size	5	Motherboard Form Factor	3
Screwdriver Tip Size	5	Wireless Capabilities	3
Roll Length	5	Recommended Room	3
Standby Power Consumption	5	Musical Instrument Type	3
Number of Power Modules	5	Pole Color	3
Antenna Type	5	Image Processor Brand	3
Optical Zoom	5	Magnet Type	3
Video Output Standard	5	Manual White Balance	3
Exposure Modes	5	Maximum Shutter Speed	3
Sports League	5	Maximum Travel Distance	3
Pet Type	5	Effective Sensor Resolution	3
ISO Range	5	Gauge	3
Streaming Services	4	Zoom Adjustment	3
Offset Distance	4	Lens Construction	3
Computer Cooling Type	4	Mounting Hardware Included	3
Base Material	4	Thickness	3
Digital Image Formats	4	White Balance Presets	3
Recommended Surface	4	Minimum Shutter Speed	3
Body Material	4	Dialer Type	3
Fill Material	4	Dialing Modes	3
Maximum Storage Temperature	4	Self-Timer Delay	3
Compatible Tape Width	4	Faceplate Style	3
Lockable	4	Maximum Image Resolution	3
Display Modes	4	Base Type	3
Maximum Operating Range	4	Focus Type	3
Minimum Storage Temperature	4	Optical Sensor Size	3
Audio Power Amplifier Class	4	Ink Color	3
Headphone Technology	4	Color Depth	3
Ergonomic	4	Pump Included	3
Number of Drawers	4	Wall Mountable	3
Number of A/V Inputs	3	Cover Material	3
Number of Lines	3	Ringer Control	3

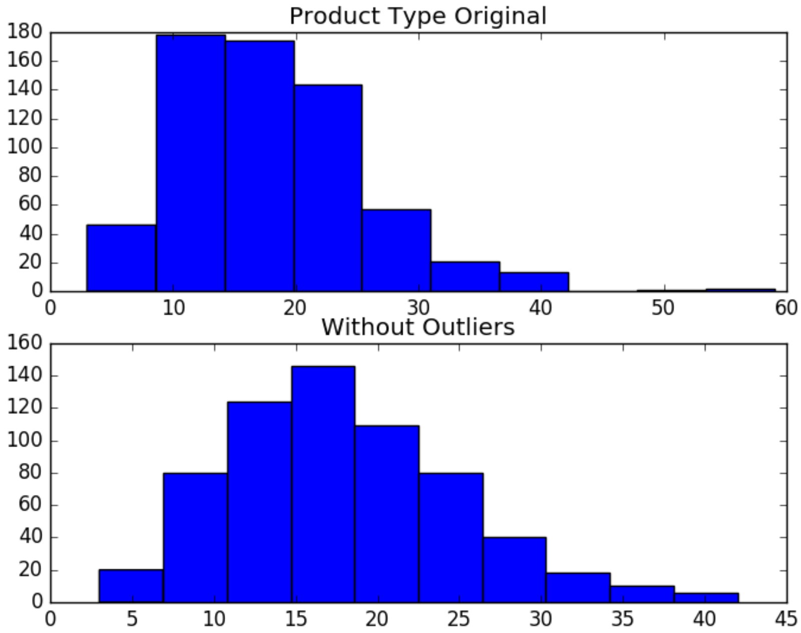
Number of Substations	3	Caller ID	2
Dialer Location	3	Charger Included	2
Number of Audio Outputs	3	Voice Control	2
Exposure Compensation	3	Printing Technology	2
Orientation	2	Vertical Viewing Angle	2
Coffee Filter Size	2	Image Resolution Yield	2
Base Color	2	Number of Controllable Devices	2
Camera Accessory Bundle Type	2	Data Storage	2
LED Indicator	2	Image Stabilization	2
Is Signal Booster	2	Stand Base Type	2
Tuner Mode	2	Sound Level	2
Maximum DVDs Held	2	Maximum View Angle	2
Voice-Activated	2	Number of Autofocus Zones	2
Pet Size	2	Effective Flash Distance	2
Web Technology Supported	2	Frame Finish	2
Wireless Network Security Protocols	2	Has Stand	2
Reading Level	2	Nominal Voltage	2
Nutritional Data Required	2	Clothing Type	2
Environmental Certifications	2	Animal Type	2
Absorbency	2	Ships in Multiple Boxes	2
Special Effects	2	Fabric Material	2
Cable Connector Type	2	Pest Type	2
Educational Focus	2	Additional Compartments	2
Hardware Lock Type	2	Microphone Technology	2
Manual Operation	2	Output Waveform	2
Number of Exposure Metering Zones	2	Minimum Focus Range	2
Has Clock	2	Microphone Output	2
Number of Speeds	2	Radio Antenna Frequency Band Type	2
Standby Time	2	Pages Per Minute	2
Aperture Range	2	Has Shoulder Strap	2
Frame Material	2	Programmable	2
Field of View Crop Factor	2	Top Material	2
Adjustable Fan	2	Display Location	2
Flash Sync Speed	2	Laptop Bag & Case Style	2
Viewfinder Type	2	Sport Type	2
Edition	2	Shooting Programs	2
Has Magnetic Shield	2	Detachable Faceplate	1
Shutter Speed Range	2	Antenna Connector Type	1
Continuous Shooting Speed	2	Warnings	1
J-Box Location	2	Number of Ringtones	1
Maximum Shooting Speed	2	Number of Cartridges	1

HD Radio	1	Adjustable Depth	1
Anti-Aging	1	Caliber	1
Drive System	1	Waterproof	1
Retractable	1	Error Correcting	1
Number of Rack Units	1	Video Recorder	1
Skin Type	1	Instructions	1
Number of Pins	1	Transmission Range	1
Alphanumeric Character	1	Compact Stereo Features	1
Designer	1	Maximum Expandable Memory	1
Fitness Goal	1	Flash Guide Number	1
Manufacturer City	1	LCD Screen Resolution	1
Partner Originated Base UPC	1	GPS Device Type	1
Message Recorder	1	Data Integrity Check Types	1
Map Datum	1	Copy Speed	1
Grip Type	1	Number of Recording Modes	1
Charging Time	1	Light Bulb Type	1
Computer Replacement Part Type	1	Number of Utilized RAM Slots	1
Computer Monitor Type	1	Number of Handsets	1
DLNA-Certified	1	Assembled Product Weight	1
Has Installed Keylock	1	Satellite-Ready	1
Coupler Type	1	Cell Phone Case Type	1
RAM Memory Speed	1	Handle Style	1
Optical Disk Drive Type	1	Bed Size	1
Circuit Breaker Type	1	Horsepower	1
Read Format	1	Shop by Personality	1
Barcode Type	1	Manufacturer Street	1
Media Included	1	Recording Mode	1
Resistance	1	Recharge Time	1
Flavor	1	Manufacturer State	1
Maximum Frequency Response	1	Depth Without Door & Handles	1
Has Lid	1	Woofer Size	1
Manufacturer Phone Number	1	BD Profiles	1
Antenna Direction	1	Remote Control Model	1
Storage Media Type	1	Rating Reason	1
Chuck Size	1	Data Usage	1
Energy Consumption Per Year	1	Network Cable Type	1
Audio Turntable Speed	1	Reverse Mode	1
Number of Recording Layers	1	Has Disc Changer	1
Decibels	1	Has Casters	1
Made From Recycled Materials	1	Cordless Phone Standard	1
Number of Tabs	1	Line Coding Format	1
Automatic Voltage Regulation	1	Speakerphone Capability	1
Number of Cameras	1	Interface Speed	1

Flash Type	1		
Framed	1		
Holding Capacity	1		
Manufacturer Web Site	1		
Cellular Network Technology	1		
Dialed Calls Memory	1		
Automatic Shutoff	1		
Audio Studio Rack Type	1		
Bicycle Frame Size	1		
Clothing Size	1		
Manufacturer Zip Code	1		
Recording Speed	1		
Number of Antennas	1		
Animal Health Concern	1		
Health Concern	1		
Industrial	1		
Latency Timing	1		
Maximum RPM	1		
Mini-Jack Adapter	1		
Scent	1		
Energy Star Version	1		

Top 10 most Frequent Attributes

1. **Product Type**
    a. Missing Values - Missing values is 0.00/34491.00 in fraction and 0.00% in percentage
    b. Technique for missing values - Not applicable as there are no missing values.
    c. Attribute classification – This attribute can be classified as **Categorical**
    d. Average/Min/Max length of values - NA
    e. Possible outliers and anomalies - As this is categorical (text) data, we detect outlier based on the Product Type's string length. Below is a histogram of the attribute value's length with and without the outliers.



For Product Type attribute, we looked for outliers based on the length of the product type. We used Median Absolute Deviation algorithm for outlier detection. We found that the Products with Product Type's length >=51 (i.e. 51, 57 and 59) are outliers.

The outlier values are as follows.

1. *Home Brewing Thermometers & Temperature Controllers,*
2. *Digital camera - SLR with Live View mode, movie recording*
3. *Cordless phone w/ call waiting caller ID & answering system*

Since these values contain valid data based on the attribute, they are not anomalies

f. <u>Attribute format</u> - The attribute values do not follow any specific format.

g. <u>Attribute synonyms</u> - Yes, there are synonyms in this attribute e.g.

   i. headset is a synonym of headsets
   ii. oscilloscopes is a synonym of telescopes
   iii. loudspeakers is a synonym of speaker

h. <u>Sprinkled values</u> - Since every row in this table contains this attribute, the value of this attribute are not sprinkled in other attribute

i. <u>Data quality problems</u> - There are a few examples for this attribute where two distinct values are clearly referring to the same thing, but have a slightly different format:

- 'Corded phones' and 'Corded phone'
- 'DVD recorder / VCR combo' and 'DVD/VCR Player Combos'
- 'Power Adapters' and 'Power adapter'

2. **Product Name**
    a. Missing Values - Missing values is 0.00/34491.00 in fraction and 0.00% in percentage
    b. Technique for missing values - Not applicable as there are no missing values.
    c. Attribute classification - This attribute cannot be classified as there are a few numeric values and majority of textual values e.g.: 0073130400012, 0073130400081 **If we were to consider these numeric values as anomalies, then our attribute can be classified as textual.
    d. Average/Min/Max length of values - NA
    e. Possible outliers and anomalies - For Product Name attribute, we looked for outliers based on the length of the product name. We used Median Absolute Deviation algorithm for outlier detection. We found that the Products with Product Name's length as >=195 are outliers. Examples are as follows
        i. *Fosmon Micro USB Slimport (MyDP) to HDMI Male to Female Cable Adapter-Connect Slimport Devices to HDTV 3D 1080p - White~FUNCTIONAL DEVICES INC / RIB TR50VA001US Class 2 Transformer, 24VAC, 50 VA, 1 PH*
        ii. *Rog Swift Pg278q 27 Lcd Monitor - 1 Ms - 2560 X 1440 - 16.7 Million Colors - 350 Nit - 100,000,000:1 - Wqhd - Displayport - Usb - Black - Energy Star 6.0, Erp, J-moss [japanese Rohs], Rohs, (pg278q)*

        Since these values contain valid data based on the attribute, they are not anomalies
    f. Attribute format - The attribute values do not follow any specific format.
    g. Attribute synonyms - There are no synonyms in these attribute values
    h. Sprinkled values - Since every row in this table contains this attribute, the value of this attribute are not sprinkled in other attribute.
    i. Data quality problems - Most product names contain information about the product's features. For example:

        'Off Lease REFURBISHED IBM Lenovo M58 C2D 3.1GHz 4GB 1TB DVD Windows 7 Home Tower Computer'

        A lot of this information might better be contained only in the product's short or long description attributes, as it does not contain any additional information about the product's name. Thus, in this case, the name could be better written as "Off Lease REFURBISHED IBM Lenovo M58". This would help remove redundancy among values of different attributes and make the product name more clear. A lot of values are, however, well structured. For example - 'Lexmark C734A1KG Black Toner Print Cartridge'.

3. **Product Segment**
    a. Missing Values - Missing values for attribute Product Name is 0.00/34491.00 in fraction and 0.00% in percentage
    b. Technique for missing values - Not applicable as there are no missing values.
    c. Attribute classification - Attribute Product Segment can be classified as Categorical.

d. Average/Min/Max length of values - NA
e. Possible outliers and anomalies - For Product Segment attribute, we looked for outliers based on the length of the product segment. We used **Median Absolute Deviation** algorithm for outlier detection. We found no outliers for this attribute.
f. Attribute format - The attribute values do not follow any specific format.
g. Attribute synonyms - Yes, there are synonyms in this attribute e.g.
    i. Default is a synonym of Everything else
h. Sprinkled values - Since every row in this table contains this attribute, the value of this attribute are not sprinkled in other attribute
i. Data quality problems - None

4. Product Long Description

a. Missing Values - Missing values is 187.00/34491.00 in fraction and 0.542170% in percentage
b. Technique for missing values - In case of missing values, for later machine learning algorithms, we will need some imputation method to figure out the missing value. Conventional imputation methods like marginal mean imputation and conditional mean imputation can not be used because each product description is different even though it belongs to the same class. Integrity constraints among attributes such as functional dependencies or application-specific “business rules” can be derived, which can be used to complete missing values. Some attributes, such as Name and Brand, often appear in the description as well, so we might construct a description of Brand + Product Name in the event that it is missing. Since these words would often be in the description of a matching product, this gives us a decent chance of identifying a match based on this attribute even when it is missing.
c. Attribute classification - Attribute Product Long Description can be classified as Textual
d. Average/Min/Max length of values - Average length of Product Long Description attribute values = 1299.13
Maximum length of Product Long Description attribute values = 3998.00
Minimum length of Product Long Description attribute values = 1.00
**We are trimming attribute values
e. Possible outliers and anomalies - For Product Long Description attribute, we looked for outliers based on the length of the product long description. We used **Median Absolute Deviation** algorithm for outlier detection. We found that the Products with product long description of length >=3969 are outliers. However, these are also descriptions and hence they are not anomalies.
f. Attribute format - The attribute values do not follow any specific format.
g. Attribute synonyms - There are no synonyms in these attribute values.
h. Sprinkled values - NA
i. Data quality problems - The product description attributes have the most text and are not categorical, and therefore are the most prone to minor data quality problems

such as spelling errors. To clean this part of the data up and increase the ability to match based on these attributes, we could run the descriptions through a spell-checker and fix any obvious mistakes. Misspellings, Illegal Values

5. **Brand**

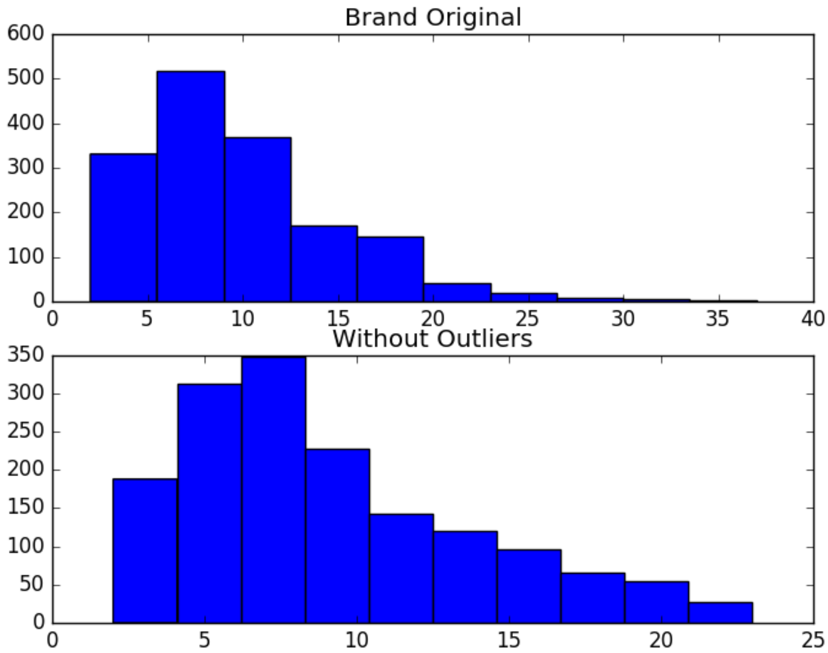
a. Missing Values - Missing values is 7030.00/34491.00 in fraction and 20.382129% in percentage

b. Technique for missing values - Product's brand is mentioned quite often in the product's short/long description and/or the name of the product. It can thus be extracted from these attributes. We could put all known brand names into a dictionary and use that to help extract the brand from these other attributes.

c. Attribute classification - Attribute Brand can be classified as Categorical

d. Average/Min/Max length of values - NA

e. Possible outliers and anomalies - For Brand attribute, we looked for outliers based on the length of the brand name. We used **Median Absolute Deviation** algorithm for outlier detection. We found that the Products with brand name of length >=24 are outliers.



Examples outliers include: *Brother International Corporat, Dr. Koffer Fine Leather Accessories*. Since these values contain valid data based on the attribute, they are not anomalies

f. Attribute format - The attribute values do not follow any specific format.

g. Attribute synonyms - Yes, there are synonyms in this attribute e.g.

i. 'Apex' and 'Apex Computer Technology'

ii. 'BIC America' and 'BIC

iii. 'Bedtime Originals by Lambs & Ivy' and 'Bedtime Originals'

h. Sprinkled values – The word Brand is frequently contained within the long description and name of the product.
For example, the product with id 17299065 contains the brand - 'Case Logic' - but the Brand attribute is absent for that product. Similarly, the product with ID 39149227 contains the brand "Asus" in the long description.

i. Data quality problems - There are some cases where there are two values for the same brand, with a minor difference in the format. The difference between what I would call a "synonym" and a "quality problem" for this attribute is that synonyms are clearly a different name for the same brand, whereas the quality issues are cases where we are one or two characters off, sometimes in error. E.g.

i. 'A. Saks' and 'A.Saks'

ii. 'Audio Technica' and 'Audio-Technica'

iii. 'Bausch & Lomb' and 'Bausch + Lomb'

6. Product Short Description

a. Missing Values - Missing values is 16514.00/34491.00 in fraction and 47.879157% in percentage

b. Technique for missing values - We can use the same heuristic as "product long description" to fill in this value. Attributes, such as Name and Brand, often appear in the description as well, so we might construct a short description of Brand + Product Name in the event that it is missing. Since these words would often be in the description of a matching product, this gives us a decent chance of identifying a match based on this attribute even when it is missing.

c. Attribute classification - Attribute short description cannot be classified as there is a numeric value in this attribute, for eg: 1406.
**If we were to consider this as an outlier, our dataset would be textual.

d. Average/Min/Max length of values - NA

e. Possible outliers and anomalies - For Product Short Description attribute, we looked for outliers based on the length of the product short description. We used **Median Absolute Deviation** algorithm for outlier detection. We found that the Products with product short description of length >=404 are outliers. Examples of outliers below:

<p><p>Transcends Wi-Fi SD card instantly adds wireless capability to your digital camera, letting you stream photos and videos to portable devices without the hassle of using cables, card readers or a computer. Take advantage of the high resolution of your digital camera and the versatility of your smartphone or tablet to share beautifully captured photos to the world as soon as you take them.</p></p>

As this is a valid description, it not an anomaly.

f. Attribute format - The attribute values do not follow any specific format.

g. Attribute synonyms - There are no synonyms in these attribute values.

h. Sprinkled values - NA

i. Data quality problems - The product description attributes have the most text and are not categorical, and therefore are the most prone to minor data quality problems such as spelling errors. To clean this data up, we could run the descriptions through a spell-checker and fix any obvious mistakes. Misspellings, Illegal Values

7. **GTIN**

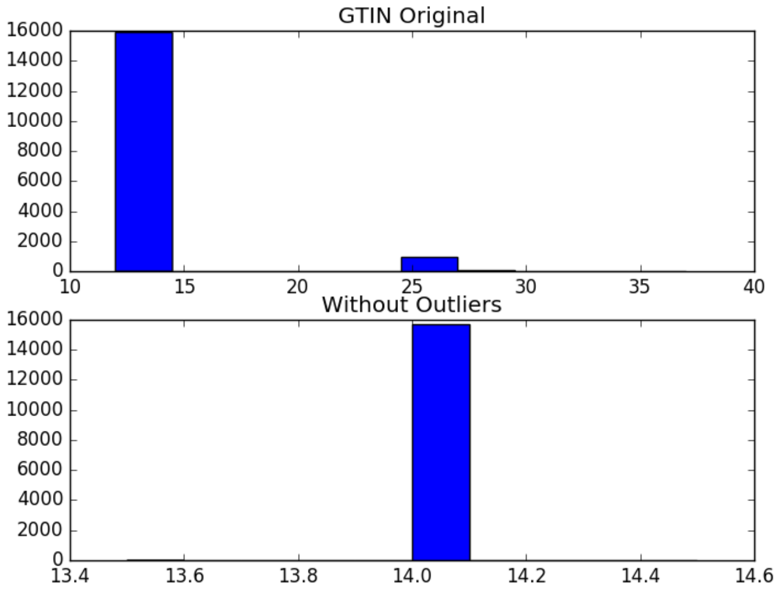
a. Missing Values - Missing values for attribute GTIN is 17497.00/34491.00 in fraction and 50.729176% in percentage.

b. Technique for missing values - **Global Trade Item Number** (GTIN) is an identifier for trade item. The uniqueness and universality of the identifier is useful in establishing which product in one database corresponds to which product in another database, especially across organizational boundaries. There is no good solution we can think of for finding the product's GTIN, unless it appears in the product's description.

c. Attribute classification - Attribute GTIN cannot be classified as there are various alphanumeric values in the text. Eg: II66OCKMTA8652.
**If we consider these as outliers, our dataset can be classified as numeric.

d. Average/Min/Max length of values - NA

e. Possible outliers and anomalies - For GTIN attribute, we looked for outliers based on the length of the GTIN attribute. We used **Median Absolute Deviation** algorithm for outlier detection. We found that the Products with GTIN value of length <> 14 are outliers.



Examples of outliers include: *GTIN_255045_00694396078318, GTIN_255045_00793442102387.*

f. Attribute format - GTIN have a universally agreed upon format of 14 or 13 or 12 or 8 digits. Various values in the dataset have the GTIN_ prefix (e.g. GTIN_255045_00793442102387) which can be removed. Some values have _ (underscores) separating various tracking numbers only one of which is a GTIN and we can standardize on what to do with the tracking numbers. This way we can arrive at a consistent image of the GTINs

g. Attribute synonyms - There are no synonyms in these attribute values

h. Sprinkled values - This attribute "sprinkled" elsewhere in other attributes.

i. Data quality problems - None

8. UPC

a. Missing Values - Missing values for attribute UPC is 17807.00/34491.00 in fraction and 51.627961% in percentage

b. Technique for missing values - Universal Product Code is again a unique code for the product. There is no good solution we can think of for finding the product's GTIN, unless it appears in the product's description.

c. Attribute classification - Attribute UPC cannot be classified as there are various alphanumeric values in the text. Eg: 60OGZV3i3850.
**If we consider these as outliers, our dataset can be classified as numeric.

d. Average/Min/Max length of values - NA

e. Possible outliers and anomalies - For UPC attribute, we looked for outliers based on the length of the UPC attribute value. We used **Median Absolute Deviation** algorithm for outlier detection. We didn't find any outliers for UPC attribute as all UPS values are of length 12.

f. Attribute format - UPC have a universally agreed upon format of 12 digits. Various values are garbled with characters, however as UPC is a representation of GTIN-12. We can use the GTIN values to standard UPC values and arrive at a standard representation of all UPC values

g. Attribute synonyms - There are no synonyms in these attribute values

h. Sprinkled values - We do not see this attribute sprinkled elsewhere.

i. Data quality problems - None

9. Country of Origin: Components

a. Missing Values - Missing values for attribute Country of Origin: Components is 19605.00/34491.00 in fraction and 56.840915% in percentage

b. Technique for missing values - We can use use certain imputation methods based on the description and product brand name. The country or origin of the product's components is often mentioned in the product's description and can be extracted from it. Besides, it should be possible to look up what the origin country of the product's brand is (possibly through a web crawler) and use that to fill in this attribute's value since it is often the case that the brand's originating country is the same as where it manufactures its components

c. Attribute classification - Attribute Country of Origin: Components can be classified as Categorical

d. Average/Min/Max length of values - NA

e. Possible outliers and anomalies - For Country of Origin: Components attribute, we looked for outliers based on the length of the Country of Origin: Components attribute value. We used **Median Absolute Deviation** algorithm for outlier detection. We didn't find any outliers for Country of Origin: Components attribute

f. Attribute format - The categorical nature of this attribute is ideal for a standard format. We can standardize to the following 4 values capturing all the cases: *USA, Imported, USA and Imported, USA or Imported*

g. Attribute synonyms - Yes, there are synonyms in this attribute e.g.

   i. US is a synonym of USA

   ii. United States is a synonym of USA

h. Sprinkled values - We do not see this attribute sprinkled elsewhere.

i. Data quality problems - It looks like there are separate categories for "USA and Imported" and "USA and/or Imported". These can be replaced by two different values - "USA and imported" and "USA or imported" which would make for better understandability.

10. Category

a. Missing Values - Missing values for attribute Category is 20922.00/34491.00 in fraction and 60.659302% in percentage

b. Technique for missing values - The category of a product is very often included in the product's name and/or long/short description and can thus be extracted from those attributes. Combination of words in the product description (for example, the mention of the terms "DDR", etc. for a RAM chip) can be used to infer the product's category. Using these as our training values, we can use established imputation methods like maximum likelihood and expectation maximum (EM) algorithm to find the missing values.

c. Attribute classification - Attribute Product Type can be classified as Categorical

d. Average/Min/Max length of values - NA

e. Possible outliers and anomalies - For Category attribute, we looked for outliers based on the length of the Category attribute value. We used **Median Absolute Deviation** algorithm for outlier detection. We found that the Products with Category value of length >=50 are outliers. Examples include: *"Math, Counting and Time Learning Toys|Geography and Social Studies Learning Toys|Language and Literacy Learning Toys"*
As this is a valid Category, it not an anomaly.

f. Attribute format - The attribute values do not follow any specific format.

g. Attribute synonyms - Yes, there are synonyms in this attribute e.g.

   i. Canes is a synonymn of Stands
   ii. Screws is a synonymn of Chisels
   iii. Puzzles is a synonymn of Joysticks
   iv. Lens is a synonymn of Lenses
   v. Beds is a synonymn of Screws

h. Sprinkled values - These values are occasionally sprinkled in other attributes like Product Name, Product Long Description and Product Short Description.
   For example, the attribute is absent for product ID 40501253 but the name of the product '3ft SVGA Super M/F Monitor Cable w/ ferrites (Gold Plated)' contains the category which is 'SVGA Monitor Cable'.
   Similarly, the product with ID 1959350#Tonzof has the product's category, which is supposed to be 'Cases', in the Product Long Description attribute. The Category attribute is missing for this product.

i. Data quality problems - This attribute seems to have some "sub-categories" appended to the end of some values, which in some cases appears to cause duplicate values where it may not be needed. E.g:

   i. 'Video Game Headsets' and 'Video Game Headsets|Headsets'
   ii. 'Laptop Sleeves|Laptop Bags|Wheeled Laptop Cases' and 'Laptop Sleeves|Wheeled Laptop Cases|Laptop Bags'

Software Tools

We have written Python script for data extraction, understanding and cleaning.

We are using the following Python packages

1. **Numpy** - The fundamental package for scientific computing in Python. We are using it for calculations like sum, median and deviation to determine outliers in attribute values.
2. **Matplotlib** - The 2D plotting library for generating histograms for determining outliers and anomalies in attribute values.
3. **NLTK** - The natural language toolkit to work with human language data. We used NLTK package to get the synonyms of attribute values.

Bonus Question

We are using the following way to determine whether the class labels are correct or incorrect.

Suppose we have three product pairs (X,Y):class1, (Y,Z):class2 and (X,Z):class3. By transitivity, if class1 is same as class2, then class3 must also be equal to class1 and class2. i.e. if class1 and class2 are MATCH then class3 should also be a MATCH. If class3 is MISMATCH, then the product pair has been misclassified.

We implemented the above technique using graph search algorithm and ran it on top of the given dataset. We did not find any mismatch using this technique.

Another way to determine whether the classification is correct or not is by checking whether all or some of the attribute values between the products are same. If the no of matching attributes between products is greater than some threshold value, we can say that the products are a MATCH.