COMP SCI 784: Stage-1

Amrita Roy Chowdhury	Ekta Sardana	Roney Michael
roychowdhur2@wisc.edu	ekta@wisc.edu	rmichael2@wisc.edu

Answers

The data consisted of 20,000 product pairs, each consisting of two products, totalling to 40,000 products. On analyzing the data, we find that of these products there were some duplicates; after deduplicating individual product records(*by finding out the products with same product ids*) we are left with **34,491 unique products** overall.

+ **The list of all attributes found in the data.**

  585 unique attributes were found in the data overall. These are listed in Appendix-A formatted as <attribute>, <number_of_products>.

+ **Analysis of top 10 most frequent attributes.**

  The top 10 most frequent attributes are stated below:

Attribute	**Number of products it appeared in**
Product Type	34491
Product Name	34491
Product Segment	34491
Product Long Description	34304
Brand	27461
Product Short Description	17977
GTIN	16994
UPC	16684
Country of Origin:Components	14886
Category	13569

- **Missing values.**

  Missing values are reported below:

Attribute	Missing Fraction	Missing Percentage
Category	20922/34491	60.66%
Country of Origin:Components	19605/34491	56.84%
UPC	17807/34491	51.63%
GTIN	17497/34491	50.73%
Product Short Description	16514/34491	47.88%
Brand	7030/34491	20.38%
Product Long Description	187/34491	0.54%
Product Name	0	0
Product Type	0	0
Product Segment	0	0

- **Filling in missing values.**

  In considering filling in missing values for attributes, the major factor to be considered is that the JSON element for a product id might already contain the requisite information but that it is only provided in the inappropriate field. E.g.: A product might not have a 'Brand' attribute, but the brand might instead be noted in the 'Product Name', 'Product Long Description', 'Product Short Description', etc.

  - Brand - A dictionary could be built consisting of all existing brand attribute values and missing values could be inferred by matching the name/description fields of the products with our dictionary entries. We could also use the value of the fairly common `Manufacturer` attribute instead as in most cases the two appear to be used interchangeably. We also observed that in the Product Long Description attribute, the brand name appears in a very specific format *<li>Brand Name : $brand_name$ <\li>   or <<li><strong>Brand Name</strong> - $brand_name$</li>* **.**We can exploit this to extract the brand from the description attributes.

- GTIN & UPC - GTIN and UPC both follow fixed formats. regular expressions could be constructed for these fields and matched in the name/description. We could now lookup these values on some standard GTIN/UPC values to check their accuracy; this however might be time consuming. Another option would be to search through other attribute values(Product Short Description, Product Long Description) for the case-insensitive strings “GTIN” and “UPC” and then attempt to match those format expressions in that vicinity.
- Country of Origin: Components - This attribute is quite meaningless for a large number of instances in the data as the product descriptions seem to indicate that the product originated from another country. We feel it would be best to try to extract this from the description fields through a dictionary approach or by exploiting the format < *Made in $country_name$* >, not only for missing values, but also even when the attribute is already present (due to its questionable nature).Some of the products have another attribute “Made in Country” which can also be used to fill in the values. Another option would be to have a dictionary of major states/cities to country and instead use a dictionary approach on the ‘Manufacturer State’ or ‘Manufacturer City’ attributes.
- Category - We could follow the same dictionary approach we used to infer missing ‘Brand’ values here. A second option is by referring to the ‘Product Type’ attribute for missing category values; from what we could understand, both attributes always have values which are extremely similar in meaning, and so we found build a map of ‘Product Type’ to ‘Category’ and use the mapped values when we find the latter to be missing.

- **Classification of attributes.**

In classifying the attributes into numeric, textual, categorical or boolean, a simple glance eliminated the possibilities of any of the 10 attributes listed above as being either numeric or boolean. Of the remaining two classes, we adopted the heuristic of classifying an attribute as categorical if the total number of unique values presented by it was less than 25% of the total number of instances it appeared in. This classification was then manually checked and found to work well.

The total number of instances in which each attribute appeared is already listed above. The number of unique values for each attribute was as below:

Attribute	Unique Count
Product Long Description	30914
Product Name	27565
GTIN	16994
UPC	16684
Product Short Description	15530
Brand	1622
Category	778
Product Type	651
Product Segment	23
Country of Origin:Components	7

The classes were identified as listed below:

Attribute	Class
Product Long Description	Textual
Product Name	Textual
GTIN	Textual
UPC	Textual
Product Short Description	Textual
Brand	Categorical
Category	Categorical
Product Type	Categorical
Product Segment	Categorical
Country of Origin:Components	Categorical

We did not find that any of the above attributes to not belong to any particular class. This was verified by manual inspection.

- **Lengths of textual attributes.**

  The average, minimal and maximal lengths of the textual attributes were:

Attribute	**Average**	**Minimal**	**Maximal**
Product Long Description	9.24	1	185
Product Name	175.08	1	2350
GTIN	1416.17	1	15686
UPC	16684	16684	16684
Product Short Description	14.26	1	342

- **Outliers and anomalies.**

  In considering outliers and anomalies, we had to take into account the type of the attributes. Histograms have been plotted for all attributes of both types, and are provided in Appendix-B in the order of their frequency in the entire data as listed previously.

  Categorical attributes being inherently unordered pose a challenge as the interpretation of a histogram is closely tied to the ordering of the bins. For the 5 categorical attributes of concern, we adopted the policy of ordering the bins by the frequency/count noted. This naturally leads to a monotonically changing histogram representation. From what we have observed, this appears to be exponentially distributed in all 5 cases, leading to evident legitimate outliers at either end of the range, but also no apparent reason to think them to be anomalies.

  For textual attributes, we used the lengths of the attribute as the bins for the histogram. This gives us a natural ordering, making it easy to identify any abnormal spikes or dips.

  - Product Name: We find two very apparent outlier spikes at lengths 119 and 120. It should also be noted that any lengths beyond 120 are few in number. We believe the sharp spikes to indicate that a great deal of data got truncated at 120 characters and this is backed up by our manual verification. The sparse tail beyond 120 instead indicates that the data consists of instances from at least 2 - and possibly more - sources, at least one source of which did not follow truncation. Due to the truncation, the data would be incomplete, and so these spikes would be anomalies.

- Product Long Description: We find numerous spikes in this case; 12-13, 26, 52, 862-863, 932, 968, 1033, 1990, 2937, 3022, 3053, 3697, 3879, etc. The spike at 12-13 is due to the SKU incorrectly being filled in for a lot of products. 26 is due to an incorrect dummy value. 52, 932 & 2937 seem to be cases of dummy XML tags with no data. 862-863, 1033, 1990, 3022 & 3879 appear to be a long generic truncated description, hence incorrect. 968 is due to a generic and incorrect description.

  All the ones mentioned above result from data corrupted in one way or another and are hence anomalies. There are very many more minor spikes here, which we are leaving out discussing here for the sake of brevity.

- Product Short Description: Some prominent spikes include 1-4, 27, 32, 75, 199 & 867-868. 1-4 is just incorrect dummy symbols. 27 & 32 are generic incorrect dummy values. The hump centered at 75 (including the spike) seems fairly legitimate in content. 199 & 867-868 seems to be cases of truncated data. 75 here is therefore just an outlier while the rest (1-4, 27, 32, 199 & 867-868) are all anomalies.

  As before there are very many more minor spikes here, which we are leaving out discussing here for the sake of brevity. Like the long description, this too has spikes too numerous to exhaustively enumerate. The take-away here is that the descriptions in their original form is quite unreliable and may hold more potential if used otherwise (e.g., in inferring missing values, etc.).

- GTIN: These are Global Trade Item Numbers which may only be 8, 12, 13 or 14 characters long. As such there are the only truly legitimate values here. There is however a minor spike in the data at 26, which is legitimate 14 digit GTINs prefixed with some other number; we consider this too to be legitimate here. As such, the only outlier here is the minor spike at 26, which is partly legitimate. The rest of the values, except 8, 12, 13 & 14, are anomalies in considering length. We do however find that GTINs in the data are often alphanumeric while they are supposed to be strictly numeric as per definition; we are unsure what exactly this represents as of now, but these may be another form of anomalies.

- UPC: All our identified UPCs are 12 in length. This seems to be in line with its definition, presenting no anomalies or outliers. As in GTIN, UPC is also supposed to be numeric by definition, but is presented here as often alphanumeric; this might be another form of anomalies.

- **Problems with format.**

  Of the 10 attributes of concern here, only 2 follow standard formats - GTIN and UPC.

  - GTIN: These may only be 8, 12, 13 or 14 characters long. There are however instances at 26, which is legitimate 14 digit GTINs prefixed with some other number. This is a format issue. Another problem here is that GTIN is supposed to be purely numeric, but the data has numerous alphanumeric instances. We will have to further investigate if this is a format (encoding) issue, but at present we see this as incorrect data instead.

  - UPC: This is supposed to be purely numeric, but the data has numerous alphanumeric instances. We will have to further investigate if this is a format (encoding) issue, but at present we see this as incorrect data instead.

- **Synonyms among attribute values.**
  - **Product Type:** We find quite a large number of synonymous sets of values and are reporting a few of them
    - Arcade Video Game Machines and Arcade Game Systems
    - RAM Memory and ram
    - Music Stands and Musical Instrument Stand
    - Wind Chimes and Wind Sculptures & Spinners
  - **Category:**The following sets of values are synonymous
    - Adapters , Adapters & Gender Chargers
    - Car Stereo Parts & Accessories, Car Speaker Accessories
    - Cleaning Products, Cleaning Tools
    - Headphone Accessories,Communication Headphone Accessories and Replacement Parts
    - Headsets , Headphones
    - Mounts and Brackets , Mounts and Holders
    - Toners and Cartridges,Printer toner, Printing ink

- **Product Segments** All the other values except for "Electronics" is anomalous for the given dataset. Hence there is no scope of finding synonyms here.

- **Product Name:** We observe that in a pair of matching products, one had extra information in the attribute "Product Name" which was missing in the other.
  eg-
  - Genius HS-02B Stereo Headset and Genius HS-02B Stereo Headset - Over-the-head
  - Dell 2150/ 2155 Compatible Cyan Yellow Magenta Toner Cartridge Set (Pack of 3) and Dell 2150/ 2155 Compatible Magenta Toner Cartridge (Pack of 3)
  - HP 60 Tri-color Original Ink Cartridge and HP Consumables CC643WN#140 HP 60 Tri-Color Ink Cartridge
- **Brand:** Attribute Brand shows different categories of synonymous values. We are reporting a few examples below.
  - Acronym and Spelled out form
    eg- HP and Hewlett Packard
  - Miscellaneous Variations in values
    - Pyle and Pyle Audio
    - StartTech and Startech.com
    - OfficeMate and Officemate International Corp
    - Brother and Brother International Corporat
    - Multi-Tech Systems and Multi-Tech
- **Country of Origin Components:** This attribute has seven different values namely
  - US
  - USA
  - United States
  - USA and/or imported
  - USA and imported
  - USA or imported
  - Imported

  Quite evidently the first three reported values (US,USA and United States) are synonymous. The next three values can be either coalesced into the single value “USA and/or imported” or treated as different values depending on the requirements of the data analyst.

- The other two attributes **GTIN** and **UPC** denote unique numbers to identify trade items. Hence case for synonymity is not applicable here.

- **Sprinkled attribute values.**
  - We find that the value for **Brand** can be found to be inside the attribute
    - **Product Name:**
      - "Product Name":["Patriot FUEL+ 6000mAh Dual-Port Rechargeable Battery"]
      - "Product Name":["ACUITY LITHONIA CS1 WWGR Power Cord F/120V High Bay Lighting"]
    - **Product Long Description:**
      - "Product Long Description":["<p>The Waber-by-Tripp Lite UL620-15 Power Strip offers a convenient method of power distribution in workbench, wallmount or floor mount applications...</p><p><b>General Information</b><ul><li>Manufacturer: Tripp Lite</li><li>Manufacturer Part Number: UL620-15</li><li>Brand Name: Tripp Lite</li>...
      - "Product Long Description":["As a high-performance hard drive alternative, Intel&reg; Solid-State Drives boost your PC to the next level in storage performance and reliability....<p><b>General Information</b><ul><li>Manufacturer: Intel Corporation</li><li>Manufacturer Part Number: SSDSC2BW120A4K5</li><li>Brand Name: Intel</li>...
    - **Country of Origin Components** can also be found in the Product Long Description and Product Short Description.
      - "Product Long Description":["<p>12-outlet surge suppressor offers protection for all electronics and is equipped with child safety covers for each outlet black housing an 8 cord with a space-saving right-angle plug and diagnostic LEDs to indicate protection and ground are present….</p><p><b>General Information</b>...<li>Country of Origin: China...
      - "Product Short Description":["Premium Compatibles HP 650A HP CE271A Cyan Laser Toner Cartridge - PCI HP CE271A HP 650A Cyan Laser Toner Cartridge for HP LaserJet Enterprise CP5520

    CP5525 CP5525CN CP5525N CP5525XH 15k Yld Made in the USA by PCI - Premium Compatibles Inc. Premium Compatibles Inc. manufactures excellent high quality Toners"]

    - Product Long Description":["FISKARS-Triple Sized 3-in-1 Embossing Corner Punch With Fiskars creativity behind each of their products you wont be disappointed This is another one of their wonderouse 3 in 1 punches with three choice designs for corners you can mix and match the corner punches and the embossing punches for nine possible effects Easy enough for scrapbooking beginners Made in Korea SKU: NMG17367"]
- **UPC/GTIN** values are found in Product Short Description and Product Long Description attributes.
    - "Product Short Description":["Premium Compatibles HP 03A C3903A 4K Laser Toner Cartridge - PCI HP 03A C3903A 4K Black LaserJet Toner Cartridge for Hewlett Packard LaserJet 6mp HP LaserJet 6P HP 6pse HP 6pxi HP 5P HP 5MC HP 5MP HP 5MV HP LaserJet 6MP replaces 6R905 LY-C3903A V703A TRH111B 02-21-0314 HEWCC3903A Made in USA UPC 845161002810"]
- **Category** and **Product Type** can also be obtained by semantic parsing of the Product Long Description , Product Short Description or Product Name.
    - "Product Long Description":["<!-- Disti Content -->The XTREME Xplosives Inner-Ear Headphones feature Extreme Deep Bass Ports and 8.5mm Neodymium driver units that deliver ultimate bass sound and rubber protectors for body durability. The Superior sound isolation and comfortable fit with ergonomic contours make these the headphones to have.<!-- END Disti Content -->"]
    - "Product Long Description":["Get the convenience of carrying along important files or footages wherever you travel, by using the SanDisk Extreme SDCZ80-064G-A75 USB Flash Drive. You can easily store and share your files with this flash drive as it has 64GB of storage space. Data transfer from machine to USB and vice-versa can happen at speed that scales up to 190MB/s. So, place an order for the SanDisk Extreme SDCZ80-064G-A75 USB Flash Drive right now!"]
    - "Product Name":["Manhattan 460729 5.0 Megapixel Webcam 500"]

        - "Product Short Description":["<p><strong>Lenovo ThinkPad T400 INTEL Core 2 Duo 2200 MHz REFURBISHED Laptop</strong></p>"

- **Data quality problems.**
    - **Incorrect Values**
        - **Product Segment.**
            - 930048-930048#Walmart.com?930048?{"Product Short Description":["Radio Raves Vol.1: Southern Gospels Top No.1 Songs Over The Past 25 Years"],"Digital Video Formats":["Cassette"],"Created By":["STAGE_PCM"],"Product Segment":["Clothing, Shoes & Accessories"]...
            - 41346986-41382013#Zoro?41346986?{"Product Name":["ALLSTAR 8833TC-OCS Radio Transmitter,3 Channel,Visor Clip G5831707"],"Product Type":["Garage Door Opener Systems & Supplies"],..."Product Segment":["Clothing, Shoes & Accessories"]...
            - 40623330-40623330#OneCall?40623330?{"Condition":["New"],"Brand":["YURBUDS"]..."Features":["Headphone Styles - In-ear, Features - Designed for Sports, Color - Gray, Refresh Rate - 20Hz - 20, 000Hz, Compatibility - iPhone (Any), Driver Unit - 15mm, Plug Type - 3.5mm, Headphone Type - Sports, Headphone Uses - Sports, Inline Microphone - Yes"],"Color":["Gray"],"Product Segment":["Jewelry, Gems & Watches"]...
    - **Conflicting Values**
        - The attributes **Country of Origin:Components** and **Product Long Description** have different values as the country of origin.
            - 41195282-41195282#TEKENVY?41195282?{"Brand":["Fellowes"]," Country of Origin: Components":["USA"],..."Product Long Description":["PlushTouch Mouse Pad with wrist support features a revolutionary FoamFusion Technology that helps relieve wrist pressure with superior comfort and softness. ...<ul><li>Country of Origin: China</li></ul></p>"],"Product Segment":["Electronics"]}?MATCH
        - Semantically attributes **Product Type** and **Category** should have similar values but for some tuples they have very different values.

        - "Product Type":["Network Switch Modules"],"Category":["Computers: Replacement Parts & Accessories"]
- **Text format problems**
    - **Product Type**
      The problem of multiple semantically equivalent variations of value is observed for the attribute Product Type as well. We are listing out a few examples:
        - Webcams and webcams
        - antenna and Antennas
        - Televisions and televisions
        - tablet_computers and Tablet Computers
        - Portable Radios and Portable radio
        - Power adapter - AC / car / USB,Power adapter - car,Power adapter,Power Adapters,Power adapter - car / USB
        - Motorcycle & ATV Accessories and motorcycle_and_atv_accessories
    - **Brands**
      The brand values suffer from different variations in the same name.
        - Case
            - intel and Intel
            - dreamGEAR and Dreamgear
            - xerox and XEROX
        - Spacing
            - Tripp Lite and Tripplite
            - Hipstreet and Hip Street
            - Tripp Lite and Tripp-Lite
            - e-replacements and ereplacements
- **Category**
  Some of the semantically redundant values noticed in this attribute are listed as follows:
    - Computers: Replacement Parts & Accessories,Computer Replacement Parts,Computer Components: Replacement Parts & Accessories
    - Cordless Phones|Additional Handsets,Cordless Phones, Corded Phones,Corded Phones|Cordless Phones
    - Computer Headsets|Headsets, Computer Headsets
    - Wheeled Laptop Cases, Wheeled Laptop Cases|Laptop Backpacks, Wheeled Laptop Cases|Laptop Bags,Laptop Bags|Wheeled Laptop

Cases|Laptop Sleeves,Laptop Bags|Wheeled Laptop Cases|Laptop Cases & Bags,Laptop Bags|Laptop Cases & Bags|Wheeled Laptop Cases,Laptop Backpacks,Laptop Bags|Wheeled Laptop Cases|Laptop Sleeves,Laptop Bags,Laptop Bags|Wheeled Laptop Cases|Laptop Sleeves,Laptop Bags|Wheeled Laptop Cases|Laptop Sleeves

- Lenses and Lens

Here we also find several values which can be coalesced into one general value or subdivided into specialised ones by removing the most generic value. For eg Cables & Connectors is the most generic category. Additionally we have specific ones like Audio Cables, Power Cord, Camera Cables & Connectors,Combination Cables|Video Cables,Computer Cables,Data Cables & Connectors,Digital Cables,KVM Cables,Microphone Cables,MP3 USB Cables, Music Instrument Cables,Networking Cables & Connectors, Printer Cables & Connectors,Telecommunication Cables etc.

- **Redundant attributes**

We thought that several attribute pairs were redundant in semantics like **Product Type** and **Category** or **Country of Origin:Components** and **Made in Country**

+ **Software tools used.**

We used a python script for analyzing the data and then Microsoft Excel to plot the histograms from the outputs. The different python libraries used were sys, json, itertools, operator and pprint. Our full script is provided in Appendix-C. Sublime Text was used for regex augmented manual verification.

+ **Bonus: Incorrect labels.**

- 42554417-41868773#TigerDirect
- 40668526-40668514#TigerDirect
- 40218730-38671698#TigerDirect
- 39552612-40985891#TigerDirect
- 33946204-21800778#UnbeatableSale.com
- 41248739-21188292#UnbeatableSale.com
- 40657847-40657825#TigerDirect
- 41251636-39805520#TigerDirect

Appendix-A

```
Product Type, 34491
Product Name, 34491
Product Segment, 34491
Product Long Description, 34304
Brand, 27461
Product Short Description, 17977
GTIN, 16994
UPC, 16684
Country of Origin: Components, 14886
Category, 13569
Manufacturer Part Number, 12725
Warranty Information, 12659
Manufacturer, 8440
Color, 8017
Actual Color, 8016
Assembled Product Length, 7893
Assembled Product Width, 7860
Assembled Product Height, 7743
Warranty Length, 3942
Condition, 2773
Composite Wood Code, 2669
E-Waste Recycling Compliance Required, 2425
Type, 2370
CPSC-Regulated Indicator, 2243
Has Mercury, 1863
Operating System, 1638
Multipack Indicator, 1539
Battery Type, 1535
Screen Size, 1466
Features, 1417
Number of Batteries, 1315
Hard Drive Capacity, 1274
Processor Type, 1235
Energy Star, 1112
RAM Memory, 1010
Memory Capacity, 917
Processor Speed, 887
Connector Type, 827
Processor Core Type, 816
California Residents Prop 65 Warning Required, 806
Electronics Certifications, 749
Release Date, 747
Material, 743
```

```
Digital Video Formats, 679
Warranty Provider, 675
Operating System Required, 675
RAM Memory Type, 652
Solid-State Drive, 652
Wireless, 590
Touchscreen, 563
Video Game Platform, 543
Monitor Type, 530
Wireless Keyboard Included, 527
Computer Mouse Included, 527
Keyboard Included, 527
Computer Mouse Type, 527
Desktop Computer Type, 520
Compatible Devices, 415
Publisher, 414
Size, 364
Interface Type, 285
ISBN-13, 247
Number of Pages, 245
Battery Life, 241
Abridged, 238
Bluetooth, 204
Batteries Included, 194
Resolution, 187
Display Resolution, 183
Recommended Use, 174
Genre, 171
Online Multiplayer Available, 167
Downloadable Content Available, 167
Cooperative Gameplay Available, 167
Subgenre, 166
Graphic Card, 165
Assembly Required, 165
Compatible Brands, 164
Depth, 163
Display Technology, 162
Volts, 160
Peripherals Required, 155
Internal/External, 152
Aspect Ratio, 151
Average Customer Rating, 145
Product Accessories Included, 132
Rack Size, 131
Energy Guide: Appliance Labeling Rule Required, 127
Number of Audio Inputs, 126
```

```
Package Quantity, 125
Widescreen, 125
Is Wireless Microphone, 122
Finish, 112
System Requirements, 112
Supported Media Formats, 107
Lifestage, 104
Form, 104
Created Date, 102
Modified Date, 102
Refresh Rate, 102
Created By, 102
Modified By, 102
Packaging Instructions, 97
Dot ID, 94
Maximum Weight, 93
HDMI Connector, 88
Mount Type, 86
Batteries Required, 79
Wireless Technology, 79
Watts, 79
Digital Audio Formats, 78
Device Type, 78
Gender, 77
Maximum Number of Offline Players, 75
Series Title, 72
Technology, 70
Internet Protocol, 68
Total Units, 68
Audio/Video Cable Type, 64
Frequency Range, 63
Earpiece Type, 62
Number of Channels, 60
Output Power, 60
Amps, 59
Processor Brand, 59
Recommended Screen Size, 58
Compatible Models, 58
Capacity, 56
Base Unit of Measure, 54
Remote Control Included, 54
Maximum Number of Online Players, 53
Age Group, 52
Record Label, 52
Audio Decoder, 48
Enclosure Color, 48
```

```
Volume Unit, 47
Maximum Recommended Age, 45
Minimum Recommended Age, 45
Has Parental Controls, 44
Cord Length, 44
Write Speed, 42
Languages, 42
Number of Speakers, 41
Maximum RAM Supported, 40
Battery Size, 40
Electronics Carrying Case Type, 40
Frequency Response Range, 40
Adjustable Tilt, 38
Analog TV Tuner, 38
Number of Poles, 38
Style, 38
Impedance, 36
Output Mode, 36
Save Big, 36
Cable Length, 34
Read Speed, 34
Mounting Pattern, 34
Response Time, 33
Shielding Material, 33
Maximum Watts Per Channel, 33
Number of Outlets, 32
Available Instant Content Sources, 32
Power Type, 32
Automatic Document Feeder, 32
Duration, 32
Connector Gender, 31
Paper Size, 31
MPAA Rating, 31
Video Game Collection, 31
Musician, 30
Power Rating, 30
Maximum Energy Surge Rating, 30
Data Transfer Rate, 30
Surge Suppression, 30
Instruction Manual Included, 29
Studio & Production Company, 29
Fastener Type, 29
Operating Power Consumption, 29
Diameter, 28
In-Line Microphone, 28
Number of Ports, 27
```

```
Age Range, 26
Automatic Two-Sided Printing, 25
DIN Size, 25
Maximum Print Resolution, 25
Minimum Operating Temperature, 25
Maximum Operating Temperature, 25
Number of Sheets, 25
Number of Discs, 24
Items Included, 24
Printing Speed, 23
Supported Networking Standards, 22
Version, 22
Instruction Manual Languages, 22
TV & Monitor Mounting Components, 22
Department of Transportation Hazard Class, 21
Headphones Included, 20
Contrast Ratio Range, 20
Maximum Load Weight, 19
Hard Drive Type, 19
Refillable, 19
Recommended Location, 18
Shipping Weight, 18
Maximum Monthly Volume, 18
Brightness, 18
On-Screen Display, 18
Actors, 17
Recording Time, 17
DVD Region Code, 17
Signal-to-Noise Ratio, 17
Sports Team, 17
Number of Megapixels, 17
Has Circuit Breaker, 17
Product Dimensions, 17
Sensitivity, 16
Internal Memory, 16
Shipping Width, 16
Diagonal Size, 16
Number of Drive Bays, 16
Shipping Height, 16
Clamping Level, 16
Shipping Depth, 16
Pattern, 16
Humidity Range, 16
Quality Tested for Walmart, 16
Headset Type, 15
Compatible Cameras, 15
```

```
Made in Country, 15
Vehicle Model, 15
Data Line Protection, 15
Color Pages Per Minute, 15
Response Bandwidth, 15
Total Harmonic Distortion, 15
Input Connector Type, 14
Speaker Driver Types, 14
Noise Filtration, 14
Apps Installed, 14
Maximum Output, 14
Number of Pieces, 14
Personalizable, 14
Title, 13
Talk Time, 13
Cell Phone Service Provider, 13
Enclosure Type, 13
Thermal Management Type, 12
Age Restriction, 12
Multifunctional, 12
Low Pass Frequency Range, 12
Theme, 12
Director, 12
Maximum Page Yield, 12
Gain Level, 12
Distance From Wall, 12
Messaging Supported, 12
HDTV, 12
Bus Speed, 11
Occasion, 11
Maximum Data Transfer Rate, 11
Software Included, 11
Volume Capacity, 11
Battery Watt Hour, 10
Number of Presets, 10
Battery Weight, 10
Microphone Included, 10
USB Version, 10
Backlight Type, 10
Viewing Angle, 10
Media Load Type, 10
Frequency, 10
Speed, 10
Character, 9
ESRB Rating, 9
Video Modes, 9
```

```
Book Type, 9
Hardware Included, 9
Has Expiration, 9
Number of Equalizers, 9
Wire Gauge, 9
Cold Crank Amp, 9
Top Mount Depth, 9
Has Cooling Fan, 9
Number in Series, 8
Printer Cartridge Type, 8
Video Streaming Quality, 8
High Pass Frequency Range, 8
Gifts by Recipient, 8
Digital Camera Type, 8
Has Warranty, 8
Surround Sound Mode, 8
Percentage of Postconsumer Content, 8
Percentage of Preconsumer Content, 8
Attachment Style, 8
Cordless, 8
Number of HDMI Connections, 8
Shape, 8
Cord Material, 7
Target Audience, 7
Recordable Media Formats, 7
Number of Line Sources, 7
Bottom Mount Depth, 7
Television Type, 7
Bass Boost Frequency, 7
Battery Backup, 7
SD Speed Class, 7
Remote Control Type, 6
USB Port, 6
Crossover Slope, 6
Compatible Cars, 6
Computer Software Type, 6
Portable, 6
Flash Modes, 6
Music Media Format, 6
Stereo Reception System, 6
HDCP Compatible, 6
Equalizer Type, 6
Number of Shelves, 6
Upscaling, 6
Has Headphone Jack, 6
Total Pixels, 5
```

```
Adjustable Height, 5
Input Signal Voltage, 5
Image Sensor, 5
Diaphragm Size, 5
Roll Length, 5
Standby Power Consumption, 5
Number of Power Modules, 5
Maximum Video Bandwidth, 5
Focal Length, 5
Video Output Standard, 5
Sports League, 5
Has Phase Shift Selector, 5
Antenna Type, 5
Pet Type, 5
Screwdriver Tip Size, 5
ISO Range, 5
Optical Zoom, 5
Exposure Modes, 5
Number of Drawers, 4
Streaming Services, 4
Offset Distance, 4
Display Modes, 4
Lockable, 4
Computer Cooling Type, 4
Base Material, 4
Ergonomic, 4
Digital Image Formats, 4
Maximum Operating Range, 4
Recommended Surface, 4
Fill Material, 4
Minimum Storage Temperature, 4
Audio Power Amplifier Class, 4
Headphone Technology, 4
Maximum Storage Temperature, 4
Compatible Tape Width, 4
Body Material, 4
Number of Substations, 3
Number of Lines, 3
Lens Construction, 3
Number of A/V Inputs, 3
Mounting Hardware Included, 3
Has Face Detection, 3
Thickness, 3
Remote Controlled, 3
Dialing Modes, 3
Image Stabilization Type, 3
```

```
Sensor Resolution, 3
Magnification, 3
White Balance Presets, 3
Minimum Shutter Speed, 3
Primary Distributor ID, 3
Dialer Type, 3
Wall Mountable, 3
Motherboard Form Factor, 3
Faceplate Style, 3
Maximum Image Resolution, 3
Recommended Room, 3
Ringer Control, 3
Zoom Adjustment, 3
Base Type, 3
Focus Type, 3
Pole Color, 3
Optical Sensor Size, 3
Musical Instrument Type, 3
Ink Color, 3
Image Processor Brand, 3
Color Depth, 3
Pump Included, 3
Cover Material, 3
Wireless Capabilities, 3
Magnet Type, 3
Manual White Balance, 3
Maximum Shutter Speed, 3
Self-Timer Delay, 3
Dialer Location, 3
Number of Audio Outputs, 3
Portable Radio Type, 3
Maximum Travel Distance, 3
Effective Sensor Resolution, 3
Gauge, 3
Exposure Compensation, 3
Coffee Filter Size, 2
Orientation, 2
Data Storage, 2
Ships in Multiple Boxes, 2
Continuous Shooting Speed, 2
Base Color, 2
J-Box Location, 2
Camera Accessory Bundle Type, 2
Edition, 2
Charger Included, 2
Is Signal Booster, 2
```

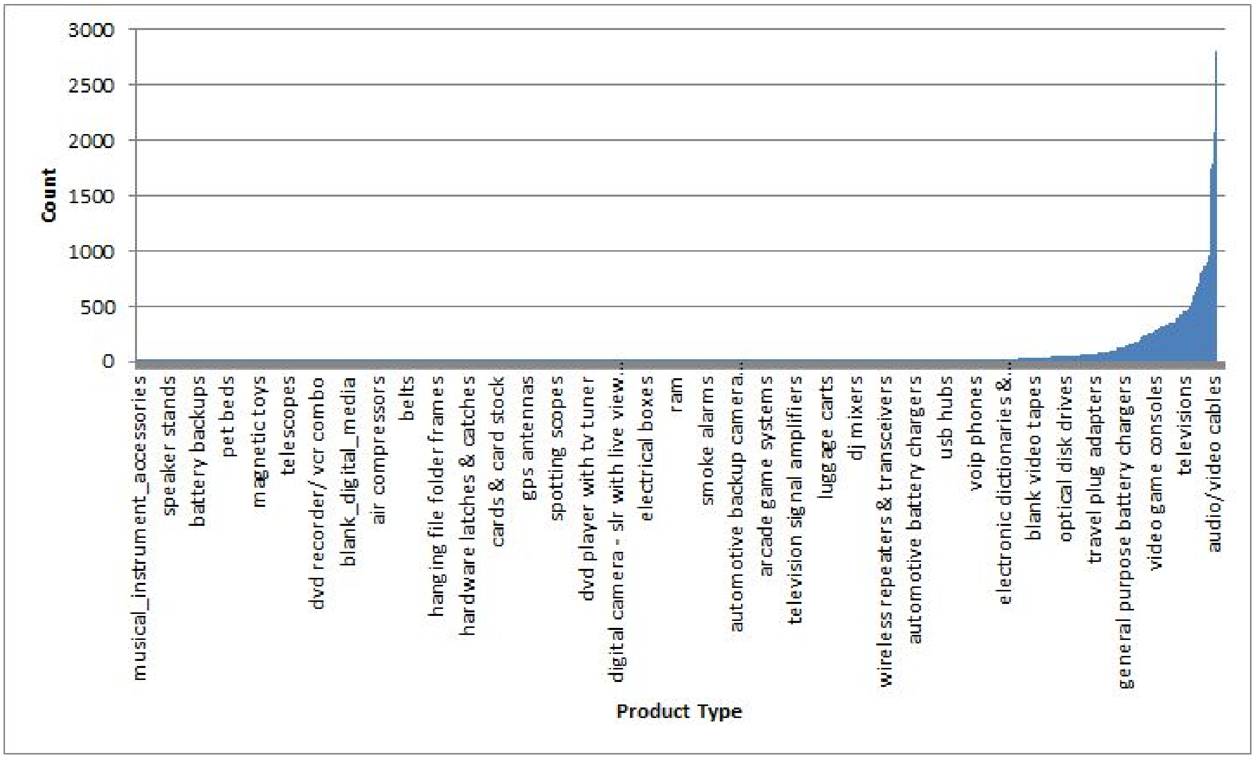
```
Voice Control, 2
Tuner Mode, 2
Maximum DVDs Held, 2
Web Technology Supported, 2
Reading Level, 2
Nutritional Data Required, 2
Environmental Certifications, 2
Printing Technology, 2
Sound Level, 2
Absorbency, 2
Vertical Viewing Angle, 2
Image Resolution Yield, 2
Number of Controllable Devices, 2
Image Stabilization, 2
Top Material, 2
Special Effects, 2
Stand Base Type, 2
Maximum View Angle, 2
Number of Autofocus Zones, 2
Effective Flash Distance, 2
Cable Connector Type, 2
Frame Finish, 2
Voice-Activated, 2
Educational Focus, 2
Hardware Lock Type, 2
Pet Size, 2
Manual Operation, 2
Caller ID, 2
Number of Exposure Metering Zones, 2
Has Clock, 2
Number of Speeds, 2
Has Stand, 2
Nominal Voltage, 2
Clothing Type, 2
Standby Time, 2
Maximum Shooting Speed, 2
Aperture Range, 2
Frame Material, 2
Fabric Material, 2
Field of View Crop Factor, 2
Adjustable Fan, 2
LED Indicator, 2
Wireless Network Security Protocols, 2
Pest Type, 2
Additional Compartments, 2
Microphone Technology, 2
```

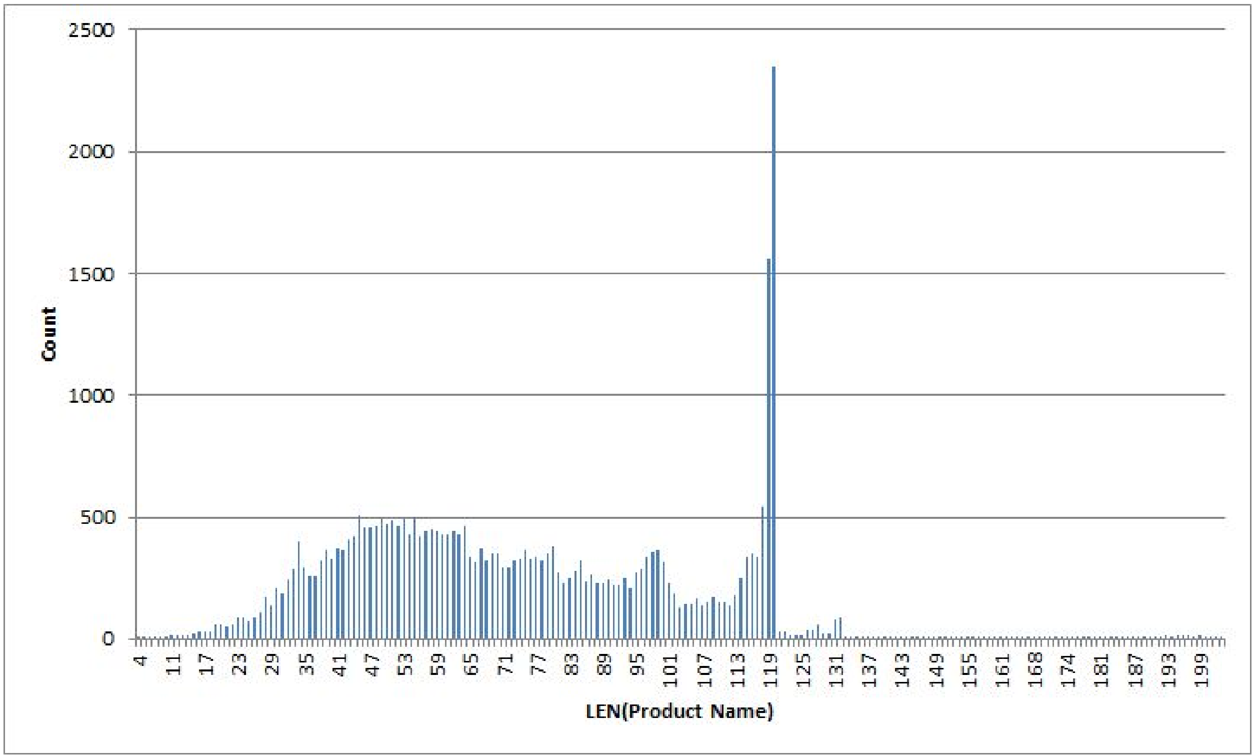
```
Output Waveform, 2
Flash Sync Speed, 2
Shooting Programs, 2
Display Location, 2
Radio Antenna Frequency Band Type, 2
Minimum Focus Range, 2
Animal Type, 2
Viewfinder Type, 2
Microphone Output, 2
Pages Per Minute, 2
Has Shoulder Strap, 2
Programmable, 2
Has Magnetic Shield, 2
Laptop Bag & Case Style, 2
Sport Type, 2
Shutter Speed Range, 2
Recording Speed, 1
Drive System, 1
Adjustable Depth, 1
GPS Device Type, 1
Detachable Faceplate, 1
Antenna Connector Type, 1
Designer, 1
Data Integrity Check Types, 1
Warnings, 1
Copy Speed, 1
Number of Recording Modes, 1
Depth Without Door & Handles, 1
Number of Ringtones, 1
Number of Cartridges, 1
Retractable, 1
Number of Utilized RAM Slots, 1
Number of Rack Units, 1
Number of Tabs, 1
Video Recorder, 1
Has Casters, 1
Skin Type, 1
Assembled Product Weight, 1
Number of Pins, 1
Satellite-Ready, 1
Cell Phone Case Type, 1
Handle Style, 1
Alphanumeric Character, 1
Bed Size, 1
Fitness Goal, 1
Partner Originated Base UPC, 1
```

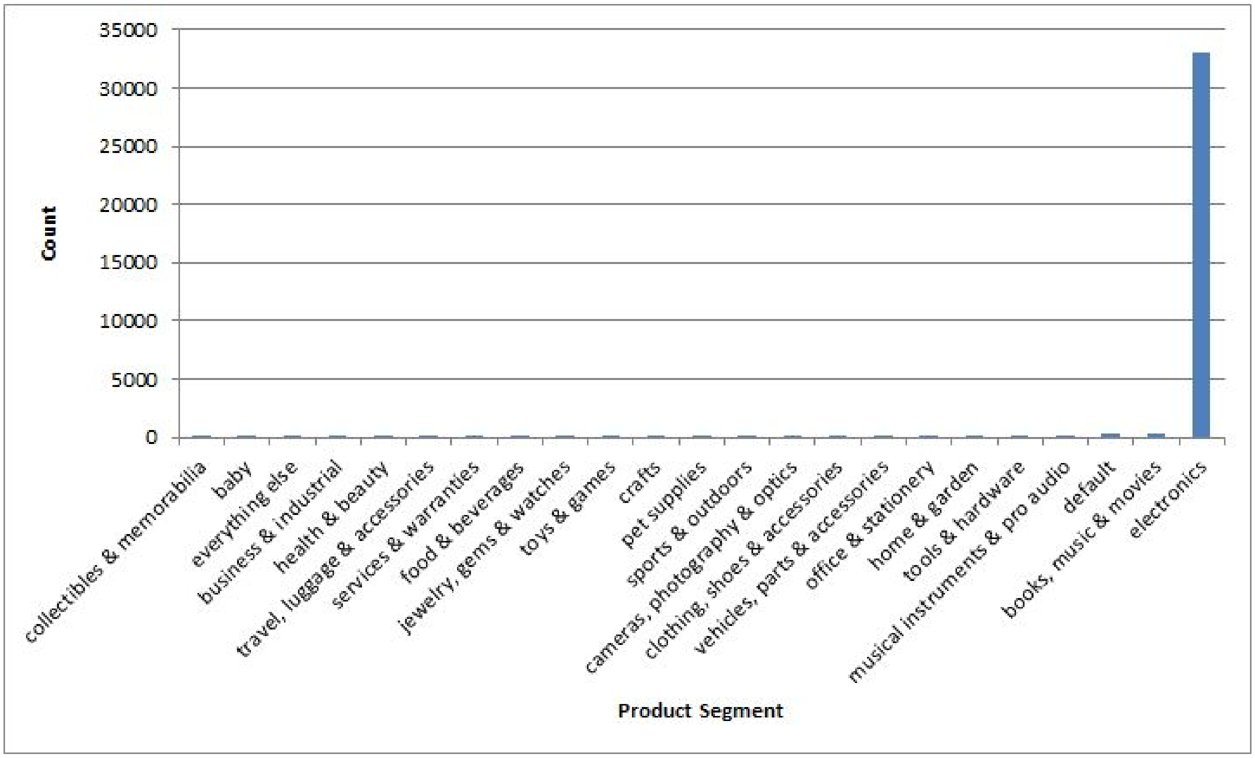
```
Horsepower, 1
Message Recorder, 1
Map Datum, 1
Shop by Personality, 1
Grip Type, 1
Woofer Size, 1
Recording Mode, 1
Recharge Time, 1
Manufacturer State, 1
Computer Replacement Part Type, 1
Manufacturer City, 1
Has Installed Keylock, 1
BD Profiles, 1
Energy Star Version, 1
Line Coding Format, 1
Rating Reason, 1
Data Usage, 1
Computer Monitor Type, 1
Chuck Size, 1
RAM Memory Speed, 1
Optical Disk Drive Type, 1
Reverse Mode, 1
Circuit Breaker Type, 1
Waterproof, 1
Storage Media Type, 1
Remote Control Model, 1
Read Format, 1
HD Radio, 1
Has Disc Changer, 1
Barcode Type, 1
Cordless Phone Standard, 1
Speakerphone Capability, 1
Flash Type, 1
Media Included, 1
Framed, 1
Coupler Type, 1
Holding Capacity, 1
Flavor, 1
Number of Recording Layers, 1
Resistance, 1
Maximum Frequency Response, 1
DLNA-Certified, 1
Anti-Aging, 1
Has Lid, 1
Cellular Network Technology, 1
Dialed Calls Memory, 1
```

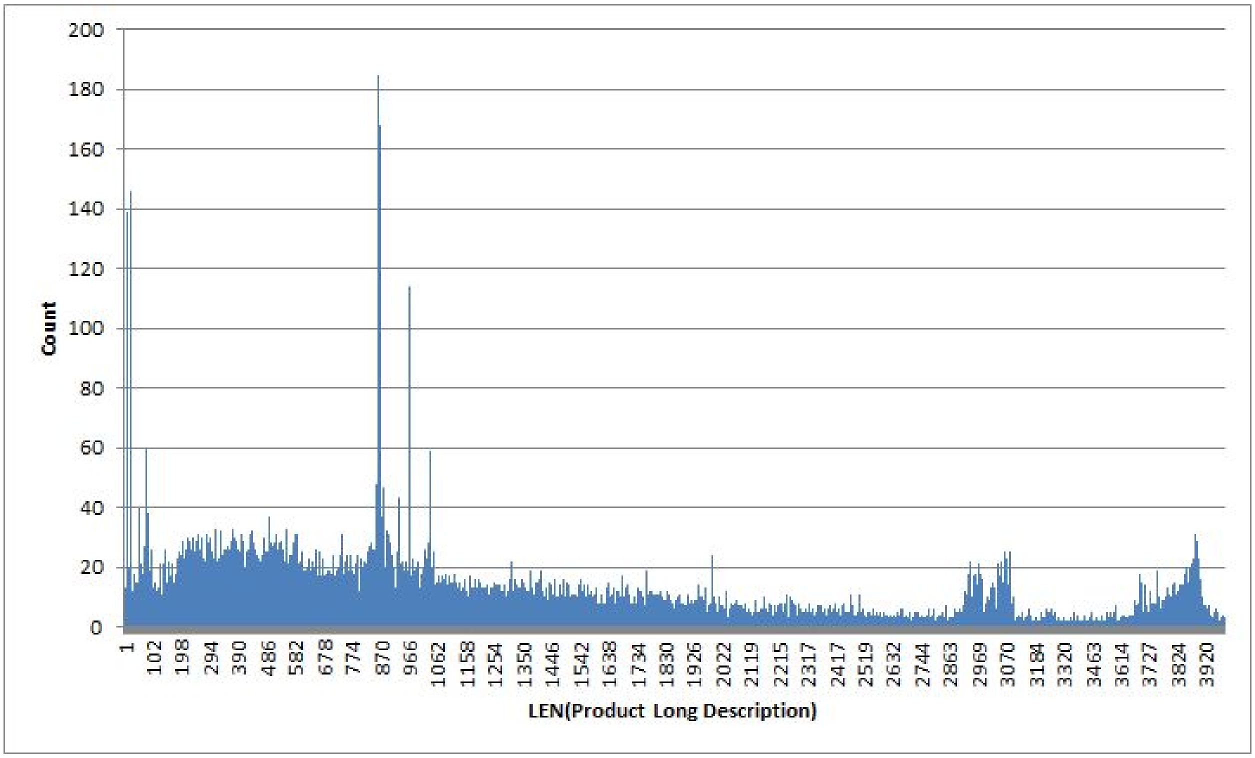
```
Light Bulb Type, 1
Charging Time, 1
Antenna Direction, 1
Number of Handsets, 1
Automatic Shutoff, 1
Audio Studio Rack Type, 1
Bicycle Frame Size, 1
Network Cable Type, 1
Energy Consumption Per Year, 1
Transmission Range, 1
Audio Turntable Speed, 1
Decibels, 1
Clothing Size, 1
Manufacturer Zip Code, 1
Industrial, 1
Number of Antennas, 1
Number of Cameras, 1
Made From Recycled Materials, 1
Manufacturer Street, 1
Health Concern, 1
Manufacturer Phone Number, 1
LCD Screen Resolution, 1
Automatic Voltage Regulation, 1
Caliber, 1
Error Correcting, 1
Animal Health Concern, 1
Latency Timing, 1
Maximum RPM, 1
Instructions, 1
Mini-Jack Adapter, 1
Compact Stereo Features, 1
Manufacturer Web Site, 1
Maximum Expandable Memory, 1
Flash Guide Number, 1
Interface Speed, 1
Scent, 1
```

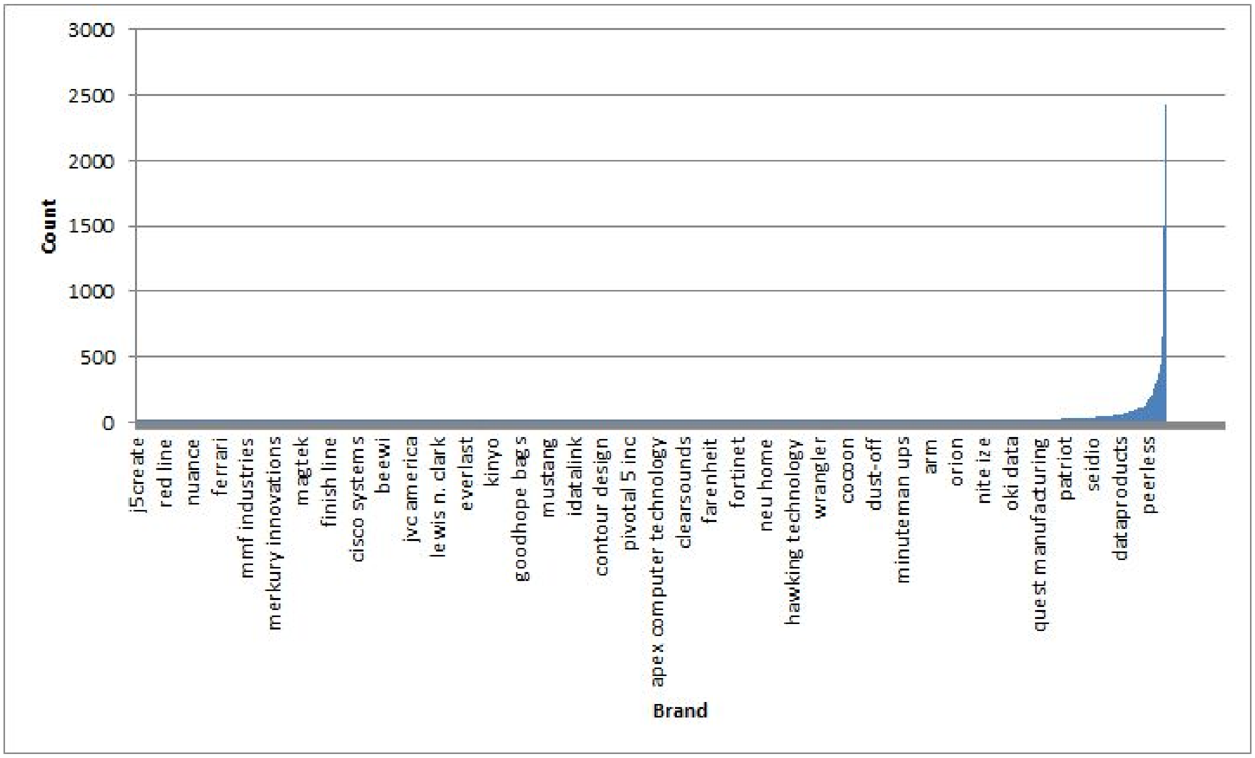
Appendix-B

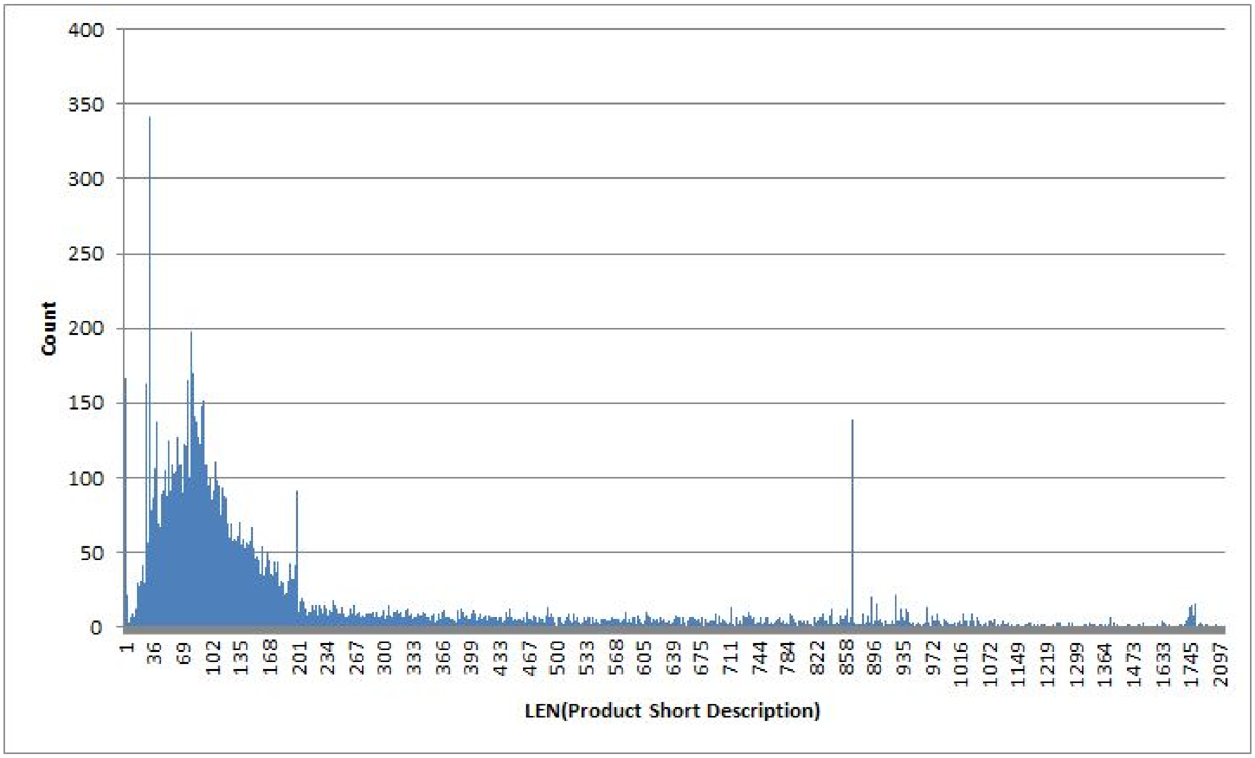

3000
2500
2000
1500
1000
500
0
Count
musical_instrument_accessories
speaker stands
battery backups
pet beds
magnetic toys
telescopes
dvd recorder/ vcr combo
blank_digital_media
air compressors
belts
hanging file folder frames
hardware latches & catches
cards & card stock
gps antennas
spotting scopes
dvd player with tv tuner
digital camera - slr with live view...
electrical boxes
ram
smoke alarms
automotive backup camera...
arcade game systems
television signal amplifiers
luggage carts
dj mixers
wireless repeaters & transceivers
automotive battery chargers
usb hubs
voip phones
electronic dictionaries &...
blank video tapes
optical disk drives
travel plug adapters
general purpose battery chargers
video game consoles
televisions
audio/video cables
Product Type

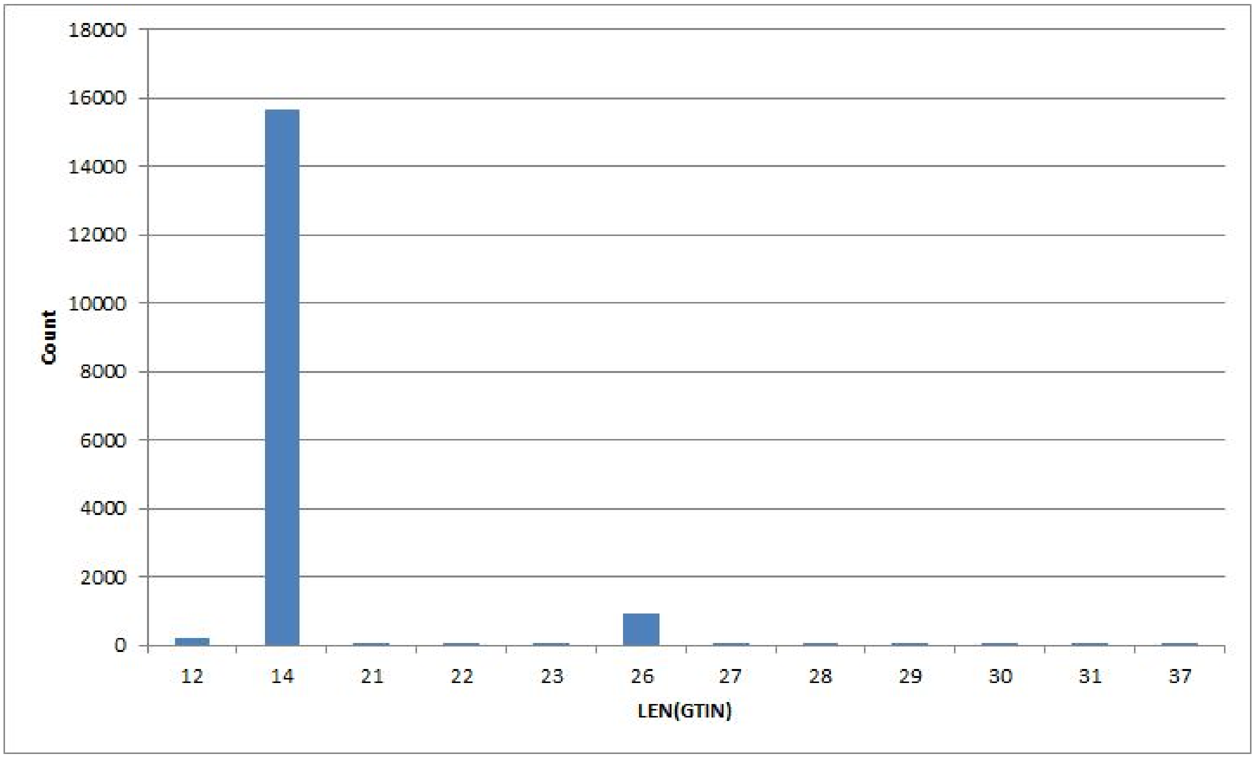

2500
2000
1500
1000
500
0
Count
4
11
17
23
29
35
41
47
53
59
65
71
77
83
89
95
101
107
113
119
125
131
137
143
149
155
161
168
174
181
187
193
199
LEN(Product Name)

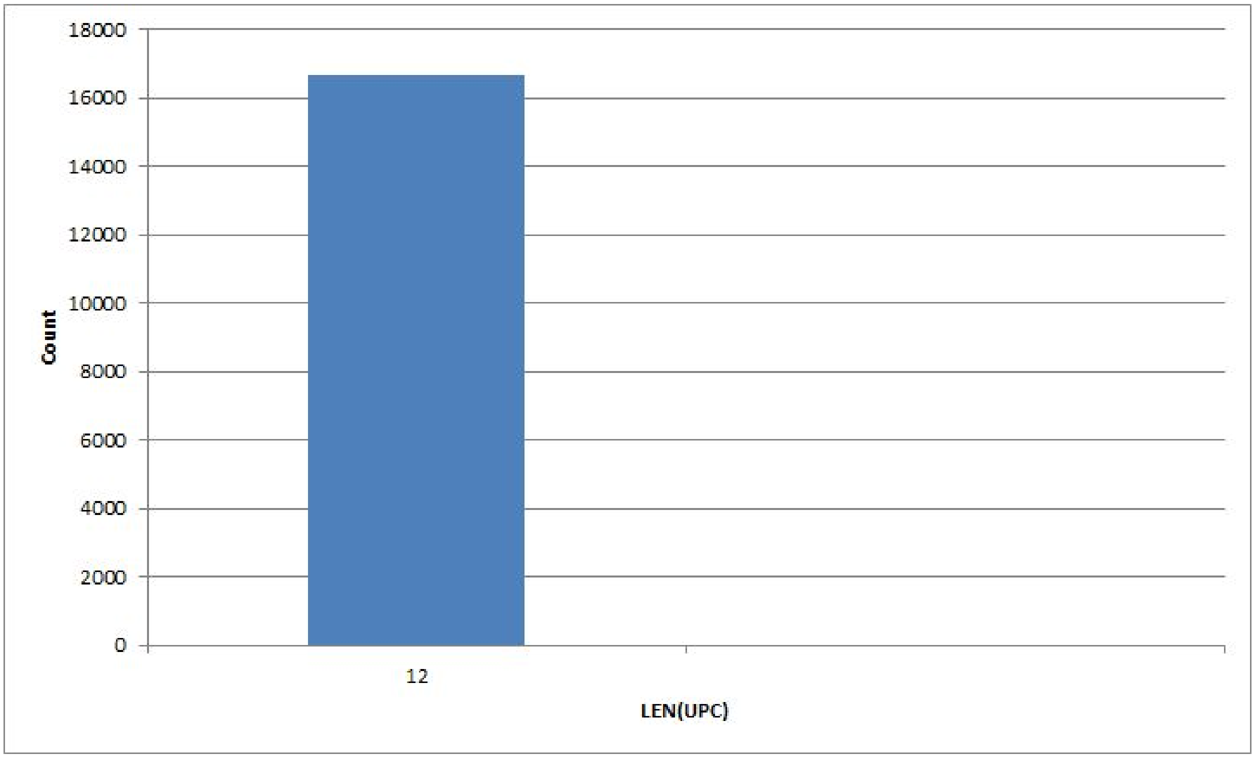

35000
30000
25000
20000
15000
10000
5000
0
Count
collectibles & memorabilia
baby
everything else
business & industrial
health & beauty
travel, luggage & accessories
services & warranties
food & beverages
jewelry, gems & watches
toys & games
crafts
pet supplies
sports & outdoors
cameras, photography & optics
clothing, shoes & accessories
vehicles, parts & accessories
office & stationery
home & garden
tools & hardware
musical instruments & pro audio
default
books, music & movies
electronics
Product Segment

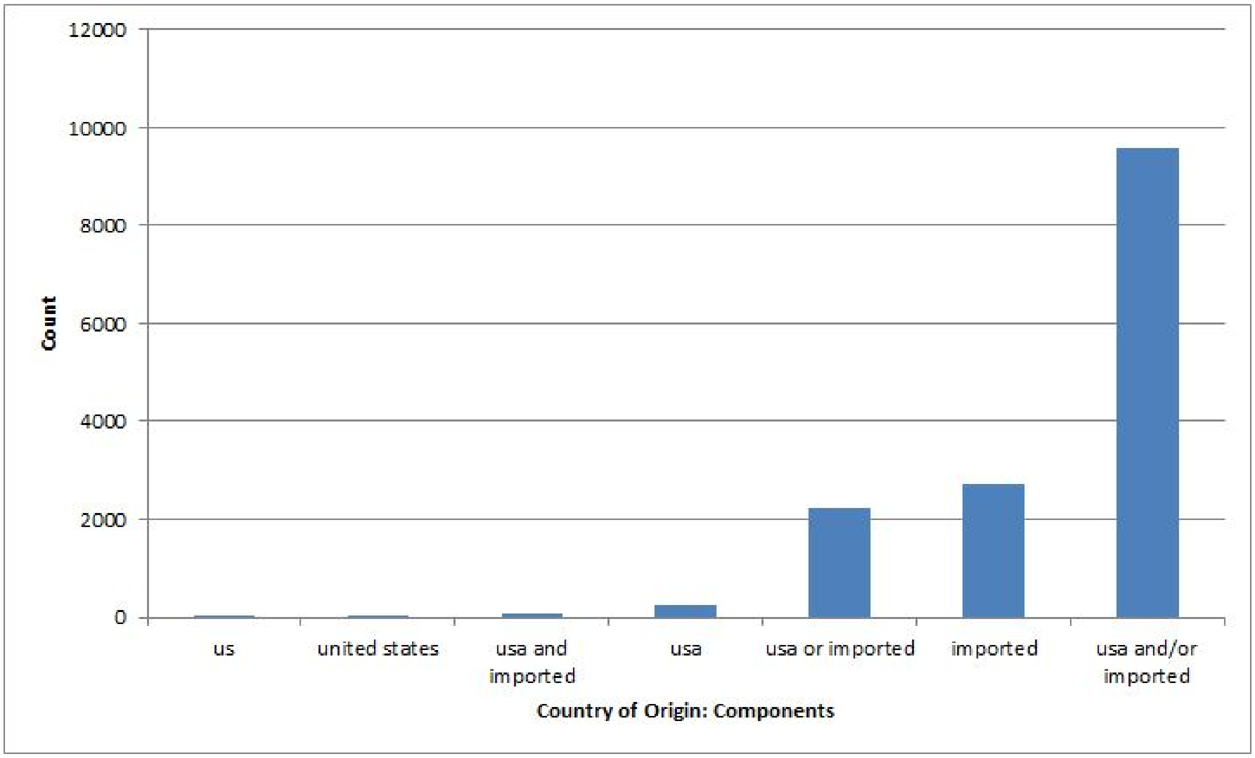

200
180
160
140
120
100
80
60
40
20
0
Count
1
102
198
294
390
486
582
678
774
870
966
1062
1158
1254
1350
1446
1542
1638
1734
1830
1926
2022
2119
2215
2317
2417
2519
2632
2744
2863
2969
3070
3184
3320
3463
3614
3727
3824
3920
LEN(Product Long Description)

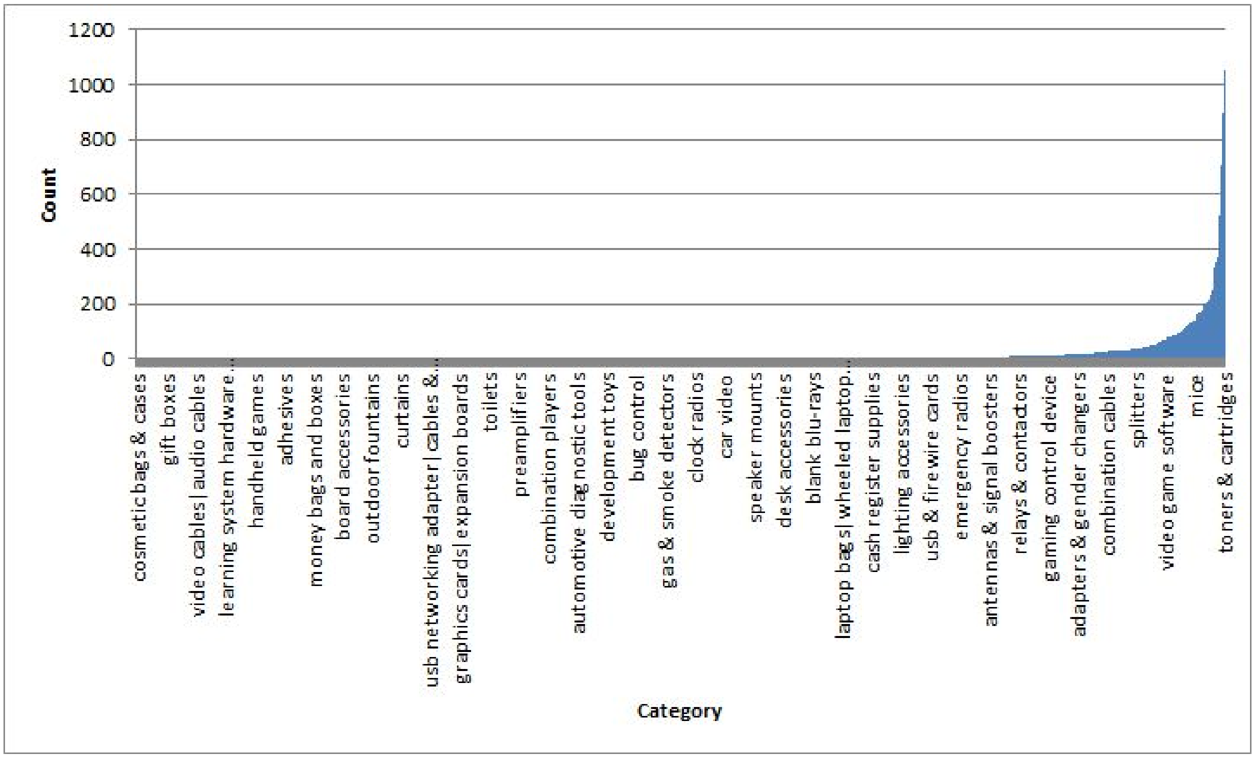

Count
3000
2500
2000
1500
1000
500
0
j5create
red line
nuance
ferrari
mmf industries
merkury innovations
magtek
finish line
cisco systems
beewi
jvc america
lewis n. clark
everlast
kinyo
goodhope bags
mustang
idatalink
contour design
pivotal 5 inc
apex computer technology
clearsounds
farenheit
fortinet
neu home
hawking technology
wrangler
cocoon
dust-off
minuteman ups
arm
orion
nite ize
oki data
quest manufacturing
patriot
seidio
dataproducts
peerless
Brand


Count
400
350
300
250
200
150
100
50
0
1
36
69
102
135
168
201
234
267
300
333
366
399
433
467
500
533
568
605
639
675
711
744
784
822
858
896
935
972
1016
1072
1149
1219
1299
1364
1473
1633
1745
2097
LEN(Product Short Description)


18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Count
12
14
21
22
23
26
27
28
29
30
31
37
LEN(GTIN)


18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Count
12
LEN(UPC)


12000
10000
8000
6000
4000
2000
0
Count
us
united states
usa and imported
usa
usa or imported
imported
usa and/or imported
Country of Origin: Components


1200
1000
800
600
400
200
0
Count
cosmetic bags & cases
gift boxes
video cables|audio cables
learning system hardware...
handheld games
adhesives
money bags and boxes
board accessories
outdoor fountains
curtains
usb networking adapter|cables &...
graphics cards|expansion boards
toilets
preamplifiers
combination players
automotive diagnostic tools
development toys
bug control
gas & smoke detectors
clock radios
car video
speaker mounts
desk accessories
blank blu-rays
laptop bags|wheeled laptop...
cash register supplies
lighting accessories
usb & firewire cards
emergency radios
antennas & signal boosters
relays & contactors
gaming control device
adapters & gender changers
combination cables
splitters
video game software
mice
toners & cartridges
Category

Appendix-C

The code we developed for the analysis is given below:

```
import sys
import json
from itertools import groupby
import operator
import pprint

if(len(sys.argv) != 2):
    print >> sys.stderr, 'Usage: python cs784stage1.py <input.txt>'

products = dict()
attributes = []
linenum = 0
with open(sys.argv[1]) as f:
    for line in f:
        linenum = linenum + 1
        line = unicode(line, errors='ignore') #For character which are not utf-8
        data = line.strip().split('?')

        pairId = data[0]
        prod1_id = data[1]
        prod1_json = json.loads(data[2])
        prod2_id = data[3]
        prod2_json = json.loads(data[4])
        label = data[5]

        if(prod1_id not in products.keys()):
            attributes = attributes + prod1_json.keys()
            products.update({prod1_id:prod1_json})
        if(prod2_id not in products.keys()):
            attributes = attributes + prod2_json.keys()
            products.update({prod2_id:prod2_json})
    f.close()

pp = pprint.PrettyPrinter(indent=4)

print 'Total number of (unique) products:', len(products)

attributes.sort()
attributeFreq = dict()
for key, group in groupby(attributes):
    attributeFreq.update({key:len(list(group))})
```

```
sortedAttributeFreq = sorted(attributeFreq.items(),
key=operator.itemgetter(1), reverse=True)

print 'Total number of (unique) attributes:', len(attributeFreq.keys())
print 'List of all attributes (<attribute>, <count>):'
pp.pprint(sorted(attributeFreq.items(), key=operator.itemgetter(1),
reverse=True))
print 'Top 10 attributes (<attribute>, <count>):'
pp.pprint(sortedAttributeFreq[0:10])

topTenAttribute = set([t[0] for t in sortedAttributeFreq[0:10]])

missing = []
attributeVals = dict()
for product in products:
    missing = missing + list(topTenAttribute -
set(products[product].keys()))
    for a in topTenAttribute:
        if(a in products[product].keys()):
            if(a not in attributeVals.keys()):
                attributeVals.update({a:[]})
            attributeVals[a] = attributeVals[a] + products[product][a]

print 'Top 10 attributeValsCount (<attribute>, <unique_val_count>):'
attributeValsCount = sorted([(a, len(set(attributeVals[a]))) for a in
attributeVals], key=operator.itemgetter(1), reverse=True)
pp.pprint(attributeValsCount)

print 'Top 10 attributeType (<attribute>, <type>):'
attributeType = []
for a in attributeValsCount:
    if len(set(attributeVals[a[0]])) < (attributeFreq[a[0]]/4):
        attributeType.append((a[0], 'categorical'))
    else:
        attributeType.append((a[0], 'textual'))
pp.pprint(attributeType)

missing.sort()
missingFreq = dict()
for key, group in groupby(missing):
    missingFreq.update({key:len(list(group))})
topMissingFreq = sorted(missingFreq.items(), key=operator.itemgetter(1),
reverse=True)
topMissingPerc = [(t[0], (100*float(t[1])/len(products))) for t in
topMissingFreq]

print 'Missing values (<attribute>, <percentage>):'
```

```
pp.pprint(topMissingPerc)

#'''
#Write all top 10 attribute values to files.
for a in attributeVals:
    fw = open('./results/attributes/' + a.replace(':', '') + '.txt', 'w+')
    for i in attributeVals[a]:
        fw.write(i.encode('utf-8') + '\n')
    fw.close()
#'''

#'''
for a, t in attributeType:
    l = attributeVals[a]
    if t is 'textual':
        l = [len(i) for i in l]
    l = dict((i, l.count(i)) for i in l)
    if t is 'textual':
        l = sorted(l.items(), key=operator.itemgetter(0))
    else:
        l = sorted(l.items(), key=operator.itemgetter(1))
    attributeVals.update({a:l})

#Write the histogram data to files.
for a, t in attributeType:
    fw = open('./results/attributes/histograms/' + a.replace(':', '') +
'.txt', 'w+')
    for i, c in attributeVals[a]:
        if t is 'textual':
            fw.write(str(i) + '\t' + str(c) + '\n')
        else:
            fw.write(i.encode('utf-8') + '\t' + str(c) + '\n')
    fw.close()
#'''
```