Assignment 1 - Data Cleanup

Table of Contents

File 1: Shoes.csv

Creation Date	03/11/2024
Version	v1.0
Reference	https://data.world/datafiniti/mens-shoe-prices

Introduction

From Datafiniti's Product Database, this file provides pricing information for over 10,000 men's shoes with 10 unique fields. This is a sample from a larger dataset.
More information about each attribute can be found at Datafiniti's Product Data Schema: https://datafiniti-api.readme.io/docs/product-data-schema.

Summary of Data Quality

Scoring Key

1 - Critical	*2 - Needs attention*	*3 - Average*	*4 - Good*	*5 - Excellent*

Category	**Description**	**Score (out of 5)**
Product and Item Segmentation	- There are a lot of vital data points that are missing from this data set. The order of the attributes (columns) could also be rearranged to improve understandability.	**2**
Duplication in Data	- Some of the data in this set is duplicated, with differences notably residing in the 'prices.*' attributes. Consider pulling this data into two separate tables to reduce load times. More info in Recommendations below. *Note: '*' notation refers to all attributes that share the common root before the asterisk*	**3**
Data Consistency	- Most data is inconsistent, and there are many erroneous data types in cells that should have one data type. For example, website links can be found in 'weight', and 'website.IDs' contains elements that are not website links.	**1**
Data Completeness	- There are large quantities of data that are missing from this data set, reducing the ability to derive meaningful patterns. - There are some attributes that are not applicable to this data of Men's shoes, such as 'vin' (Vehicle Identification Number) and 'flavor'.	**1**
Marketing Content	- The vast majority of internationally identifiable information, including ASIN, UPC, and EAN, are missing from this data set, which hinders the reliability of tracing the product. - Most brand names and product names are intact, but there are other marketable aspects of the shoes (such as color) that are missing.	**2**

Implementation Recommendations

General Notes

-

Product and Item Segmentation

- Move all ‘prices.*’ columns to the far right to improve understandability.
- Move column ‘name’ to immediately after ‘brand’ for clarity.
- Move column ‘weight’ to immediately after ‘dimension’ for clarity.
- Move column ‘sourceURLs’ to immediately before ‘imageURLs’ for clarity.
- Move columns ‘asins’ (Amazon Standard Identification Number), ‘ean’ (European Article Number), and ‘upc’ (Universal Product Code) in succession to each other immediately after ‘keys’ for clarity.

Duplication in Data

- Some of the rows of data in this set are duplicated, with the primary changes appearing in the attributes prices.*’ Consider pulling this data into two tables:
    - Table 1 would have all attributes, except the ‘prices.*’ attributes.
    - Table 2 would have only ‘id’ in addition to all ‘prices.*’ attributes.
        - By pulling the data in this way, you reduce load times when managing each separate table while still keeping the ability to reference ‘prices.*’ data with their respective products (using the common attribute ‘id’ as the link)

Data Consistency

- The columns ‘ean’ and ‘upc’ have some cells in scientific notation. Convert all of these to plain numbers for unit consistency and to improve understandability.
- Reformat the data in the following columns in the following ways:
    - Reformat and convert ‘weight’ values so that units are consistent. I recommend using the units of ounces (“oz”). This way, if you need to convert the data to pounds, you can utilize a simple unit conversion.
- There are multiple instances of information residing in columns they should not reside in. Some examples include the following:
    - There are misplaced numerical values (perhaps ‘upc’ values) placed in the ‘weight’ column, in addition to websites listed as the values in the ‘weight’ column.
    - The ‘website.IDs’ column has non-website entries in it.
    - Looking through the source URLs, data is being pulled from Ebay, Walmart, Sears, Amazon, among other sites. It is possible that when the data from these sources were gathered, the extracted data contained a different number of columns than the destination table’s columns, resulting in data being “shifted” over by one or more cells. Investigate how the data for

these entries are being pulled from their respective databases. Examples of these “shifted” rows can be found at the following:

- Row 18,000
- Row 18,001

Data Completeness

- The column ‘count’ contains no values. Delete this column to improve table load times.
- The columns ‘flavor’ (used for consumables), ‘prices.flavor’ (used for consumables), ‘isbn’ (used for books), and ‘vin’ (used for vehicles), are irrelevant to the product in question (men’s shoes). Delete these columns.
- The AW column, which contains no heading, has 3 values of “4.0 lbs”. Move these to the ‘weight’ column (AV column).
- Roughly 84% of ‘asin’ values are blank (~ 16,000 of 19,000 rows).
- Roughly 42% of ‘upc’ values are blank (~ 8,000 of 19,000 rows).
- Roughly 28% of ‘ean’ values are blank (~ 5,500 of 19,000 rows).
- Roughly 42% of ‘colors’ values are blank (~ 8,000 of 19,000 rows).
- Nearly all products have a recorded minimum value (‘prices.amountMin’), with the exception of the presence of erroneous data.
- Nearly all products have a recorded maximum value (‘prices.amountMax’), with the exception of the presence of erroneous data.

Marketing Content

- All product names (‘name’) are intact.
- A vast majority of ASIN, UPC, and EAN numbers are missing.
- Most of the product brand names (‘brand’) are intact.
- Nearly half of the product colors are missing.
    - Can you cross-reference data from other sources to receive this information, such as from Amazon using ASIN numbers?
- Nearly all products have a recorded minimum value and maximum value, with the exception of the presence of erroneous data in these columns.

File 2: Electronics.csv

Creation Date	03/11/2024
Version	v1.0
Reference	https://data.world/datafiniti/electronic-products-and-pricing-data

Introduction

From Datafiniti's Product Database, this file provides pricing information for over 7,000 electronics products with 10 unique fields. This is a sample from a larger dataset. More information about each attribute can be found at Datafiniti's Product Data Schema: https://developer.datafiniti.co/docs/product-data-schema.

Summary of Data Quality

Scoring Key

1 - Critical	*2 - Needs attention*	*3 - Average*	*4 - Good*	*5 - Excellent*

Category	**Description**	**Score (out of 5)**
Product and Item Segmentation	- While there is valuable data in this source, including official product brands, names, color, and dimensions, this aggregate data could be rearranged to improve understandability.	**4**
Duplication in Data	- Most of the data in this table is duplicated, with the only difference showing in the 'reviews.*' attributes. Consider pulling this data into two separate tables to reduce load times. More info in Recommendations below. *Note: '*' notation refers to all attributes that share the common root before the asterisk*	**2**
Data Consistency	- Most data in this set with units have inconsistent unit types. Example: the 'weight' attribute has units: "ounces", "oz", "pounds", "lbs", etc. Convert all numbers in this column into one unit type.	**2**
Data Completeness	- Most data is present, and there is likely enough to be able to cross-reference after having pulled this data, but there are still missing chunks of data, even in seemingly duplicated rows. Detailed information on this is in Recommendations below.	**3**
Marketing Content	- There are multiple sources of lookup IDs in this dataset, including ASIN, UPC, and EAN. For readability, move all of these columns to be beside each other. - Most other marketing data is intact, notably the product's name.	**4**

Implementation Recommendations

General Notes

- The column ‘primaryCategories’ is always “Electronics.” Delete this column to reduce table load times.

Product and Item Segmentation

- Move all ‘reviews.*’ columns to the far right, to increase understandability.
- Move column ‘name’ to immediately after ‘brand’ for clarity.
- Move column ‘weight’ to immediately after ‘dimension’ for clarity.
- Move column ‘sourceURLs’ to immediately before ‘imageURLs’ for clarity.
- Move columns ‘asins’ (Amazon Standard Identification Number), ‘ean’ (European Article Number), and ‘upc’ (Universal Product Code) in succession to each other immediately after ‘keys’ for clarity.

Duplication in Data

- Some rows are nearly exact duplicates of each other, with the only notable difference in the column ‘reviews.dateSeen’. Is this column needed? If not, consider deleting it and merging the remaining duplicate rows.
    - Example: id: “AVwvGPRyU2_QcyX9R3FW” with reviews.title: “BSOD with Dell's WIFI Drivers”
- Most of the rows of data in this set are duplicated, with the primary changes appearing in the attributes ‘reviews.*’ Consider pulling this data into two tables:
    - Table 1 would have all attributes, except the ‘reviews.*’ attributes.
    - Table 2 would have only ‘id’ in addition to all ‘reviews.*’ attributes.
        - By pulling the data in this way, you reduce load times when managing each separate table while still keeping the ability to reference ‘reviews.*’ data with their respective products (using the common attribute ‘id’ as the link)
- There are duplicates between some data in the column ‘ean’ and the first element of ‘keys’.

Data Consistency

- Consider merging columns ‘manufacturer’ and ‘manufacturerNumber’.
    - Before merging, note that some cells of ‘manufacturer’ are blank, while all cells of ‘manufacturerNumber’ are populated. Try cross-referencing with other data sets if possible to fill in the empty cells of ‘manufacturer’.
- Reformat the data in the following columns in the following ways:
    - Reformat ‘dateAdded’ and ‘dateUpdated’ time values to a more readable format.
    - Reformat ‘reviews.date’ and ‘reviews.dateSeen’ time values to a more readable format.

    - Reformat ‘ean’ and ‘upc’ numbers to remove the Scientific Notation, which can cause confusion.
    - Reformat and convert ‘weight’ values so that units are consistent. I recommend using the units of ounces (“oz”). This way, if you need to convert the data to pounds, you can utilize a simple unit conversion.
- Some brands have incorrect names. For example, “Dell” vs “Delll”.

Data Completeness

- There are blanks in some cells of the following columns: [asins, brand, colors, dimension, ean, manufacturer, reviews.date, reviews.doRecommend, reviews.numHelpful, reviews.text, reviews.title]
    - Notably, some rows that are exact duplicates otherwise, have some data missing. For example:
        - id: “AVwvGPRyU2_QcyX9R3FW” has multiple missing ‘asin’ numbers, despite being duplicate rows otherwise. Fill in these missing rows with the proper ‘asin’ numbers from other entries
    - Consider replacing blanks in ‘reviews.numHelpful’ with 0, assuming that blanks would mean that no one has stated that the review was helpful.
- Replace blanks in ‘reviews.rating’ with “unknown” or some other known null value to indicate that the review was not completed or valid.

Marketing Content

- All product names are intact.
- Most ASIN, UPC, and EAN numbers are intact.
- Most brand names are intact.
- There are noticeable details missing in some product ‘color’ attributes.
    - Can you cross-reference data from other sources to receive this information, such as from Amazon using ASIN numbers?