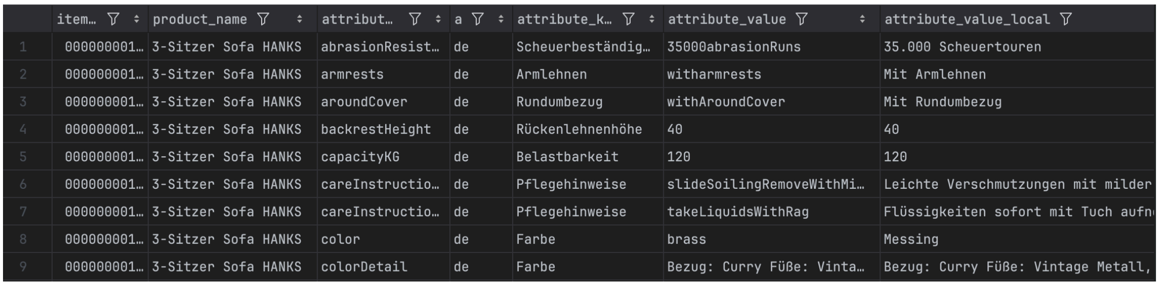
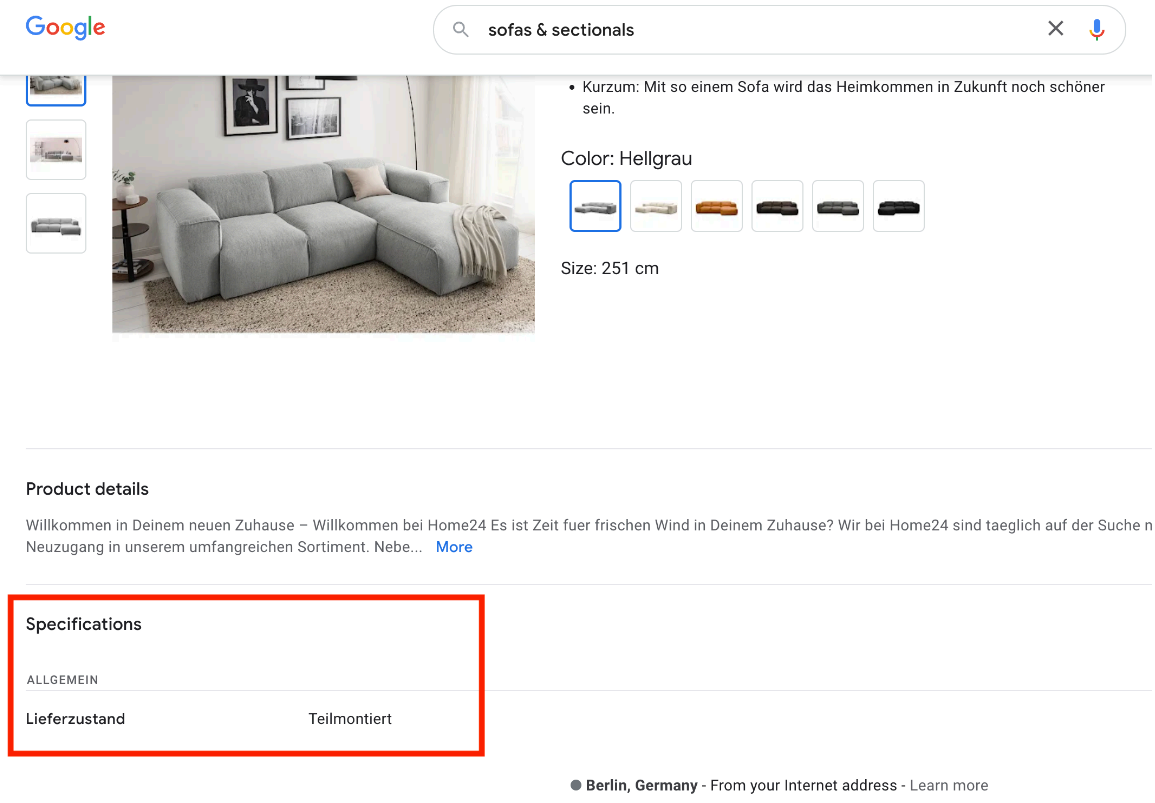
Non-Technical Summary

The Goal: Messy Data → Structured Ads

Our raw product information is stored across many files and isn't always perfectly organized for marketing. The goal was to sift through all this data, clean it up, and intelligently combine the most important product features into a single, detailed description that meets **Google Shopping's best practices**. This makes our ads more informative and appealing to potential buyers.



	item…	product_name	attribut…	a	attribute_k…	attribute_value	attribute_value_local
1	000000001…	3-Sitzer Sofa HANKS	abrasionResist…	de	Scheuerbeständig…	35000abrasionRuns	35.000 Scheuertouren
2	000000001…	3-Sitzer Sofa HANKS	armrests	de	Armlehnen	witharmrests	Mit Armlehnen
3	000000001…	3-Sitzer Sofa HANKS	aroundCover	de	Rundumbezug	withAroundCover	Mit Rundumbezug
4	000000001…	3-Sitzer Sofa HANKS	backrestHeight	de	Rückenlehnenhöhe	40	40
5	000000001…	3-Sitzer Sofa HANKS	capacityKG	de	Belastbarkeit	120	120
6	000000001…	3-Sitzer Sofa HANKS	careInstructio…	de	Pflegehinweise	slideSoilingRemoveWithMi…	Leichte Verschmutzungen mit milder
7	000000001…	3-Sitzer Sofa HANKS	careInstructio…	de	Pflegehinweise	takeLiquidsWithRag	Flüssigkeiten sofort mit Tuch aufn
8	000000001…	3-Sitzer Sofa HANKS	color	de	Farbe	brass	Messing
9	000000001…	3-Sitzer Sofa HANKS	colorDetail	de	Farbe	Bezug: Curry Füße: Vinta…	Bezug: Curry Füße: Vintage Metall,



- Kurzum: Mit so einem Sofa wird das Heimkommen in Zukunft noch schöner sein.

Color: Hellgrau

Size: 251 cm

Product details

Willkommen in Deinem neuen Zuhause – Willkommen bei Home24 Es ist Zeit fuer frischen Wind in Deinem Zuhause? Wir bei Home24 sind taeglich auf der Suche n Neuzugang in unserem umfangreichen Sortiment. Nebe... More

Specifications

ALLGEMEIN

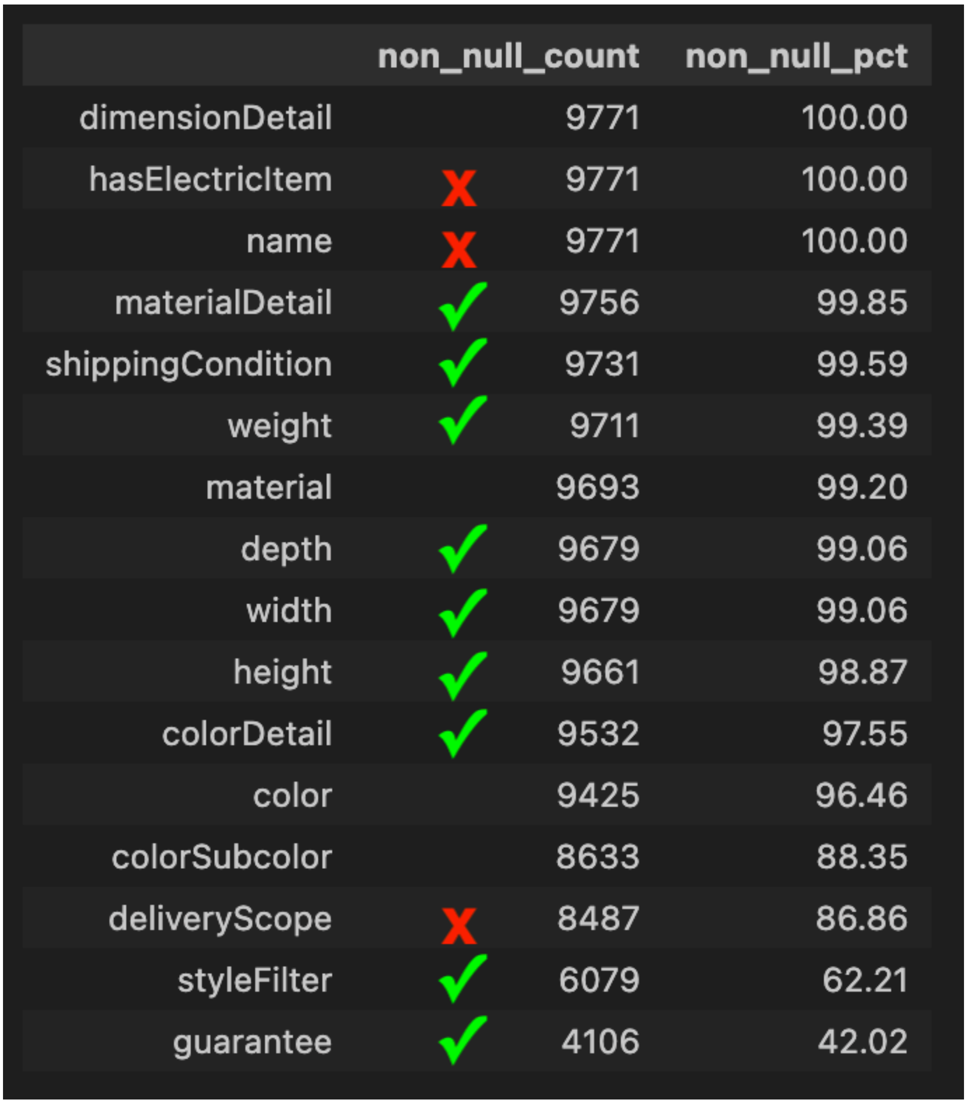
Lieferzustand	Teilmontiert

● **Berlin, Germany** - From your Internet address - Learn more

What I chose & Why

Firstly, I shortlisted 16 attributes that have less than 60% missing values. With the consideration of scalability and efficiency, I chose to prioritize the attributes that cover most of the products to display on Google Shopping Ads. Selecting 'niche' attributes only benefit a small portion of products listing while missing the majority.

Secondly, I chose 9 attributes marked in the screenshot below: **x** stands for hard exclusion, ✔ stands for the attributes I choose.



		non_null_count	non_null_pct
dimensionDetail		9771	100.00
hasElectricItem	X	9771	100.00
name	X	9771	100.00
materialDetail	✓	9756	99.85
shippingCondition	✓	9731	99.59
weight	✓	9711	99.39
material		9693	99.20
depth	✓	9679	99.06
width	✓	9679	99.06
height	✓	9661	98.87
colorDetail	✓	9532	97.55
color		9425	96.46
colorSubcolor		8633	88.35
deliveryScope	X	8487	86.86
styleFilter	✓	6079	62.21
guarantee	✓	4106	42.02

I chose `materialDetail` , `colorDetail` , `weight` , `depth` , `width` , `height` , `shippingCondition` , `styleFiler` , `guarantee` .

Here's a breakdown of the 16 attributes, the rationale behind my choice.

My selection was driven by one question: ‘**What information helps a customer make a confident purchase?’**

- **Look & Feel (Material & Color):** These are vital for online furniture shopping.

  Among the candidate attributes, `material` and `materialDetail` can overlap to each other. The same to `color`, `colorDetail` , and `colorSubcolor`.

  By taking a closer look at the data, I found that `colorDetail` usually provides more information and sometimes the combination of the other two. `colorDetail` provides richer, more descriptive detail, such as "Tischplatte: Akazie Braun Gestell: Vintage Metall".

  The same to `materialDetail`; it provides more detailed information. Instead of just ‘wood,’ our ads can now specify details like “Material: Bezug: Stoff”. This gives customers a much clearer picture of the product. Therefore, I decide to move on with `colorDetail` and `materialDetail`.

- **Practical Dimensions (Width, Height, Depth, Weight): ‘**Will it fit?’ is a key question for buyers. By providing cleaned, accurate dimensions, we answer this question upfront, reducing customer uncertainty and potential returns. Due to the time limit and complexity of `dimensionDetail`, I choose not to breakdown down the complex details which might also include numbers, units, and many more details; instead I choose to use existing individual `depth`, `height`, `weight`, `width` attributes to provide basic, satisfying information.

- **Style & Function (Style, Shipping Condition):** Attributes like Stil: Modern (Style: Modern) help customers find products that match their taste. Information about the `Lieferzustand` (`shippingCondition`, e.g., "Assembled") manages expectations about assembly. This is one of the most important consideration when shopping furniture online.

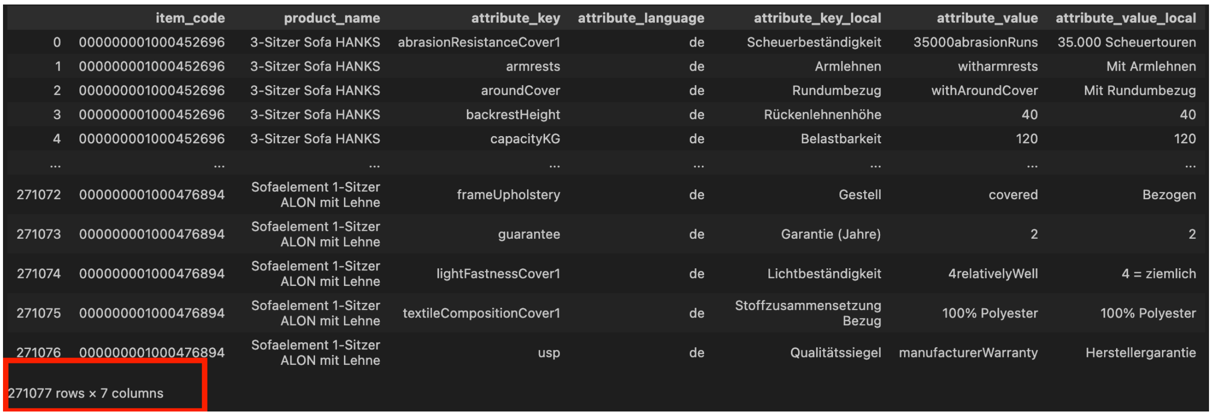
- **Trust & Assurance (Guarantee):** Including the warranty period (`Garantie (Jahre):2`) is a small detail that builds significant trust and highlights the quality of our products.

What to exclude: `hasElectricItem` seems like an attribute for internal audit, perhaps used to flag products for special handling. `deliveryScope` also seems like an attribute for internal logistics when packing for shipping. They might sometimes provide valuable information, but they are not key to customer decision making. As for `name`, it’s totally unstructured, condensed data more suitable for title; listing it in the section of specifications will only add noise.

My Problem-Solving Process

Step 1 - Take a look at data

First, I used **Python** code and a data-handling tool called **pandas**, to put all the data into a big ‘spreadsheet’. This is how it looks like:



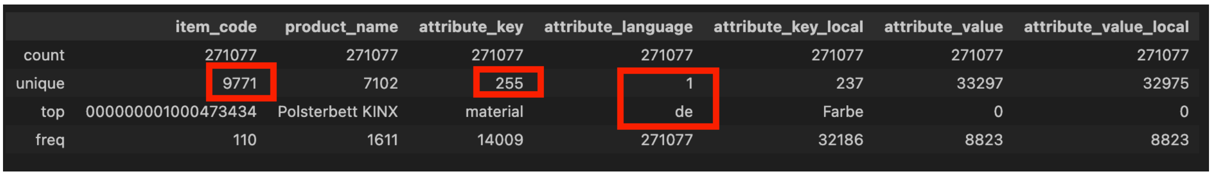
	item_code	product_name	attribute_key	attribute_language	attribute_key_local	attribute_value	attribute_value_local
0	000000001000452696	3-Sitzer Sofa HANKS	abrasionResistanceCover1	de	Scheuerbeständigkeit	35000abrasionRuns	35.000 Scheuertouren
1	000000001000452696	3-Sitzer Sofa HANKS	armrests	de	Armlehnen	witharmrests	Mit Armlehnen
2	000000001000452696	3-Sitzer Sofa HANKS	aroundCover	de	Rundumbezug	withAroundCover	Mit Rundumbezug
3	000000001000452696	3-Sitzer Sofa HANKS	backrestHeight	de	Rückenlehnenhöhe	40	40
4	000000001000452696	3-Sitzer Sofa HANKS	capacityKG	de	Belastbarkeit	120	120
...	...	...	...	...	...	...	...
271072	000000001000476894	Sofaelement 1-Sitzer ALON mit Lehne	frameUpholstery	de	Gestell	covered	Bezogen
271073	000000001000476894	Sofaelement 1-Sitzer ALON mit Lehne	guarantee	de	Garantie (Jahre)	2	2
271074	000000001000476894	Sofaelement 1-Sitzer ALON mit Lehne	lightFastnessCover1	de	Lichtbeständigkeit	4relativelyWell	4 = ziemlich
271075	000000001000476894	Sofaelement 1-Sitzer ALON mit Lehne	textileCompositionCover1	de	Stoffzusammensetzung Bezug	100% Polyester	100% Polyester
271076	000000001000476894	Sofaelement 1-Sitzer ALON mit Lehne	usp	de	Qualitätssiegel	manufacturerWarranty	Herstellergarantie

271077 rows × 7 columns

This ‘spreadsheet’ has 7 columns and over 200,000 rows. Each row stands for a piece of detail of a single product, which can be potentially selected to list on Google Shopping Ads product feed.

Step 2 - Statistics and Making Decision

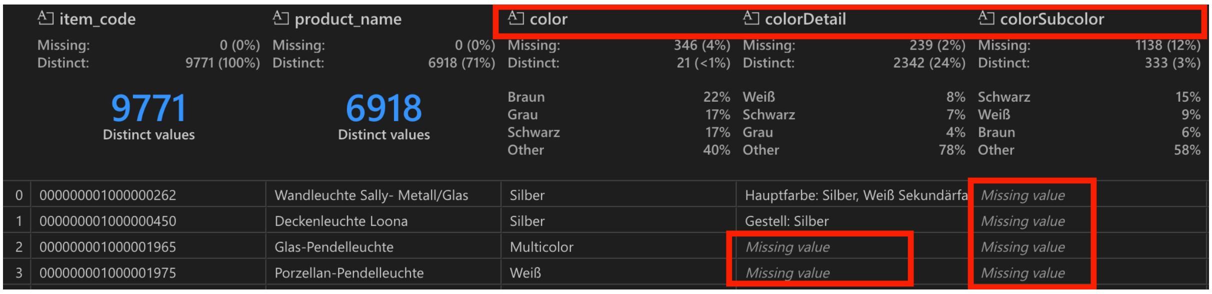
To quickly get overall information of the data, I used a built-in tool to generate a summary of the ‘spreadsheet’. It shows that the spreadsheet contains:



	item_code	product_name	attribute_key	attribute_language	attribute_key_local	attribute_value	attribute_value_local
count	271077	271077	271077	271077	271077	271077	271077
unique	9771	7102	255	1	237	33297	32975
top	000000001000473434	Polsterbett KINX	material	de	Farbe	0	0
freq	110	1611	14009	271077	32186	8823	8823

- 9771 products
- 255 attributes in the pool to choose from
- German as the only local language

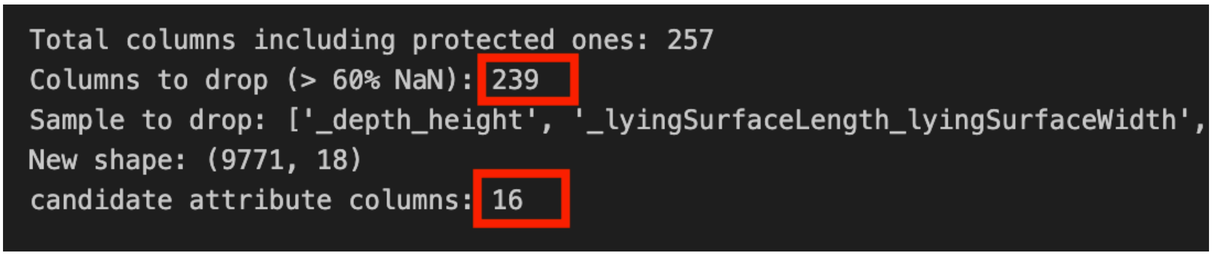
For a better view of all the attributes, I transformed the original ‘spreadheet’ by taking 255 attributes as columns and each product as a single row:



	item_code	product_name	color	colorDetail	colorSubcolor
	Missing: 0 (0%) Distinct: 9771 (100%)	Missing: 0 (0%) Distinct: 6918 (71%)	Missing: 346 (4%) Distinct: 21 (<1%)	Missing: 239 (2%) Distinct: 2342 (24%)	Missing: 1138 (12%) Distinct: 333 (3%)
	9771 Distinct values	6918 Distinct values	Braun 22% Grau 17% Schwarz 17% Other 40%	Weiß 8% Schwarz 7% Grau 4% Other 78%	Schwarz 15% Weiß 9% Braun 6% Other 58%
0	000000001000000262	Wandleuchte Sally- Metall/Glas	Silber	Hauptfarbe: Silber, Weiß Sekundärfa	*Missing value*
1	000000001000000450	Deckenleuchte Loona	Silber	Gestell: Silber	*Missing value*
2	000000001000001965	Glas-Pendelleuchte	Multicolor	*Missing value*	*Missing value*
3	000000001000001975	Porzellan-Pendelleuchte	Weiß	*Missing value*	*Missing value*

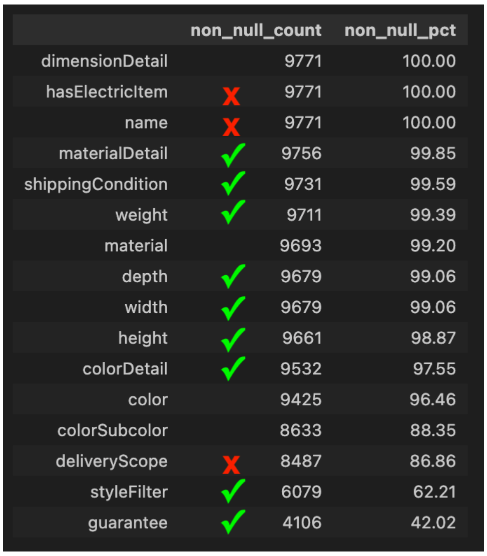
The new spreadsheet clearly shows missing values. In other words, some products are missing some attribute information. For example, Porzellan-Pendelleuchte only has a color 'Weiß', without further colorDetail provided.

I dropped the attribute columns that have over 60% missing values. If most of the products don't have the attributes, I exclude them.



```
Total columns including protected ones: 257
Columns to drop (> 60% NaN): 239
Sample to drop: ['_depth_height', '_lyingSurfaceLength_lyingSurfaceWidth',
New shape: (9771, 18)
candidate attribute columns: 16
```

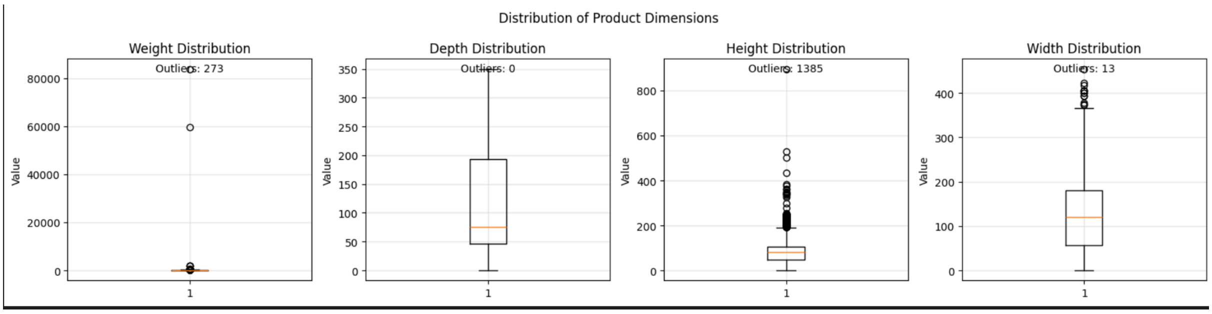
After the operation, my record tells that 239 attributes are dropped because of low coverage and only 16 left. These 16 attributes form the candidate list where I finally chose from. Here is the statistics on 16 candidate attributes' coverage, ranked from high to low.



		non_null_count	non_null_pct
dimensionDetail		9771	100.00
hasElectricItem	✗	9771	100.00
name	✗	9771	100.00
materialDetail	✓	9756	99.85
shippingCondition	✓	9731	99.59
weight	✓	9711	99.39
material		9693	99.20
depth	✓	9679	99.06
width	✓	9679	99.06
height	✓	9661	98.87
colorDetail	✓	9532	97.55
color		9425	96.46
colorSubcolor		8633	88.35
deliveryScope	✗	8487	86.86
styleFilter	✓	6079	62.21
guarantee	✓	4106	42.02

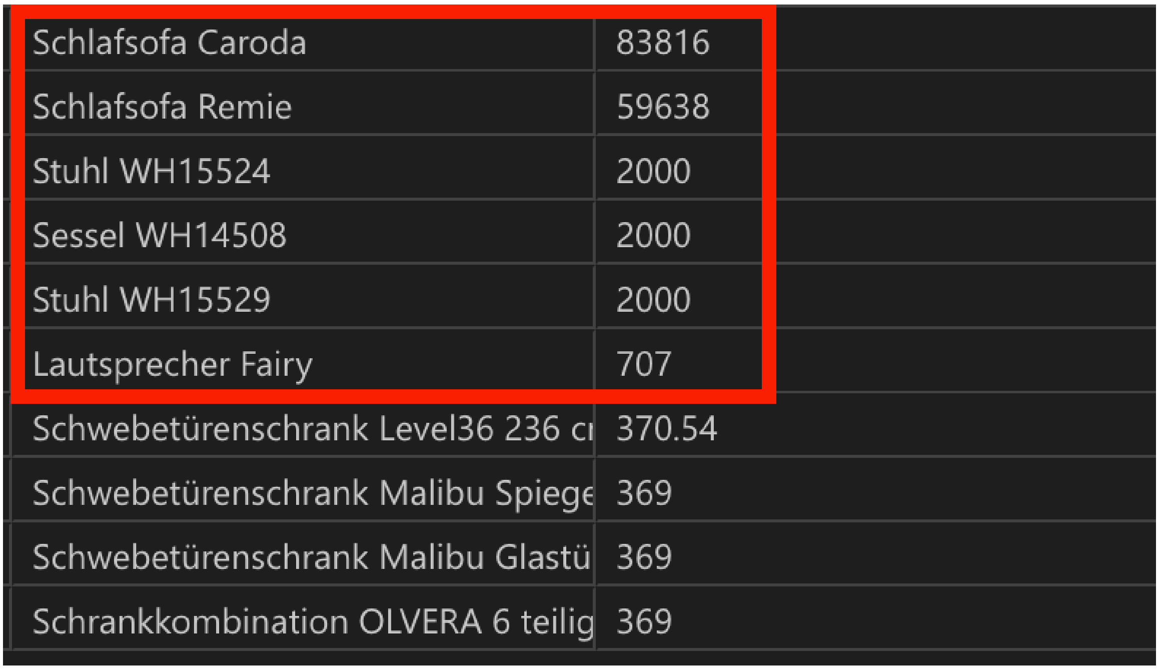
Step 3 - Cleaning and Tidying Up

Raw data can have errors, especially with numbers. I made plots below on data of weight, depth, height, width using **matplotlib**, a plotting tool on Python.



These plots shows outliers. For example, we can tell from the plot that some pieces of furniture are weighed as about 60000 kg or over 80000 kg, which are obviously unrealistic.

I listed the top 10 heavies items, and found that the errors are probably due to the missing decimal. A sofa should be around 80 kg at most. So I set a threshold: any items weighing over 600 kg will automatically get a decimal back.

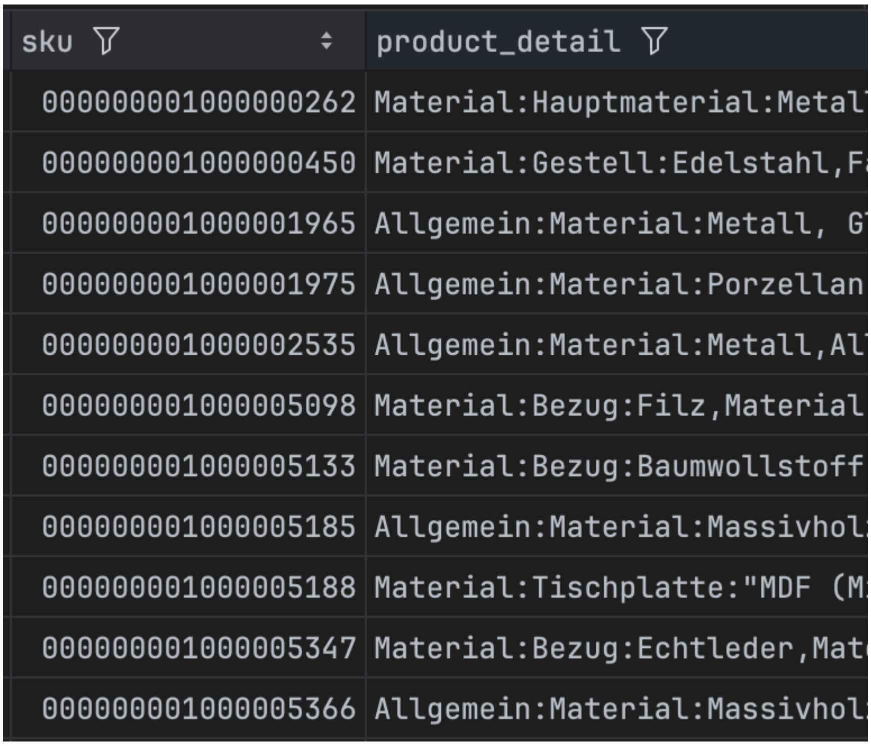


Schlafsofa Caroda	83816
Schlafsofa Remie	59638
Stuhl WH15524	2000
Sessel WH14508	2000
Stuhl WH15529	2000
Lautsprecher Fairy	707
Schwebetürenschrank Level36 236 cı	370.54
Schwebetürenschrank Malibu Spiege	369
Schwebetürenschrank Malibu Glastü	369
Schrankkombination OLVERA 6 teilig	369

In the same way, any furniture with height over 300 cm, which exceeds average ceiling height, can hardly realistically fit into a normal room. Thus, I set threshold for height (300 cm), depth (200 cm), width (500 cm), and cleaned the data. These are rough estimations and can be adjusted based on product categories.

Step 4 - Formatting to comply to Google

Finally, I broke down the attributes that have combined information, such as `colorDetail` and `materialDetail`, and combined all the chosen attributes in local language into the special format required by Google (`section_name:attribute_name:attribute_value`). Then I export it as a clean, ready-to-use **CSV file** named **product_detail.csv**. This file can be directly uploaded to the Google Merchant Center as supplemental feed:



sku	product_detail
0000000001000000262	Material:Hauptmaterial:Metal
0000000001000000450	Material:Gestell:Edelstahl,F
0000000001000001965	Allgemein:Material:Metall, G
0000000001000001975	Allgemein:Material:Porzellan
0000000001000002535	Allgemein:Material:Metall,Al
0000000001000005098	Material:Bezug:Filz,Material
0000000001000005133	Material:Bezug:Baumwollstoff
0000000001000005185	Allgemein:Material:Massivhol
0000000001000005188	Material:Tischplatte:"MDF (M
0000000001000005347	Material:Bezug:Echtleder,Mat
0000000001000005366	Allgemein:Material:Massivhol

Result

Example output on Google Shopping Ads:

Specifications

Material

Hauptmaterial	Metall
Sekundärmaterial	Glas

Farbe

Hauptfarbe	Silber, Weiß
Sekundärfarbe	Weiß

Allgemein

Gewicht	1.27 kg
Tiefe	15 cm
Breite	82 cm
Höhe	45 cm
Stil	Modern
Lieferzustand	Montiert
Garantie (Jahre)	2

Improvement

If I am given more time, I will make the following improvement.

- **Product category-based Data Cleaning and Attribute Selection:**

  Incorrect data can erode trust for customers. Right now, the rules for spotting data errors and attribute selection are general.

  We could make them much smarter by using the **product category**. For example, a normal weight for a small speaker (0.7kg) is very different from that of a big sofa (80kg). In the same sense, a ‘degree of hardness’ can be a very important information to buy a pillow, but might be irrelevant for most of the tables.

To improve, I would integrate AI to be an assistant who classifies all the products into different categories, then let it define the proper range of dimension of each category. For each category, I would select the most frequent attributes and rank them. In this case, data cleaning will be more target to ensure the data to present to customer is more accurate and to ensure the data follow the best practice on Google Merchant Center.

- **AI-Powered Language Check:** While our data is intended to be in German, mistakes can happen. We could use **AI to automatically verify the language** in the detail. This would catch any accidental English phrases that slip in, ensuring a consistent and professional experience for our customers.
- **Uncovering More Attributes:** We currently only present the features of our products which show up more frequently to keep things tidy. However, some of these rare features could be valuable for customers and powerful for marketing. Particularly, `dimensionDetail` has valuable details , which can be broken down to present customers with more detailed dimension.