In [1]:

```
import pandas as pd

pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
```

In [2]:

```
# load data
#df = pd.read_csv("raw_product_data.csv", encoding='utf-8')
df = pd.read_csv("raw_product_data.lower.csv", encoding='utf-8')
#df = pd.read_csv("analysis/x.csv")
left_width = max(len(str(key)) for key in df.columns)
```

In [3]:

```
print('\nNumber of unique values for each column:\n')
for col in df.columns:
    count = len(df.groupby(col).count())
    print(f"{col:<{left_width}} {str(count):>10}")
```

```
Number of unique values for each column:

variant_id      1090941
product_id       284650
size_label         1348
product_name     407770
brand               731
color              1278
age_group             2
gender                3
size_type             1
product_type       5872
```

In [4]:

```
# All lines with the same variant_id contain the same product_id
df[['variant_id', 'product_id']].groupby('variant_id').filter(lambda x: x['product_id'].nunique() != 1)
```

Out[4]:

variant_id	product_id

In [5]:

```
# Two rows with the same product_id may have different brand
rows1 = df[['product_id', 'brand']].groupby('product_id').filter(lambda x: x['brand'].nunique() != 1)
```

```
rows1
```

Out[5]:

	product_id	brand
11	13817074	ami paris
25	13817144	ami paris
40	15134938	ami
47	15134577	ami paris
76	14371972	ami
...	...	...
3097499	14942233	ami paris
3097520	17276624	adamo
3097541	14942752	ami paris
3097548	13810414	ami paris
3097580	17482131	alice+olivia

160682 rows × 2 columns

In [6]:

```
# Number of affected rows
rows1[['product_id']].groupby('product_id').count()
```

Out[6]:

product_id
10510878
10687443
10687444
10945322
10948012
...
19036472
19038124
19107205
19107269
19485859

7910 rows × 0 columns

In [7]:

```
df[df['product_id'] == 19038124]
# Output the list of brands that are affected
x = rows1.groupby('product_id')['brand'].agg(list)
#for product_id, brands in x.items():
#    print(f"Product ID: {product_id}")
#    u_brands = set(brands)
#    for brand in u_brands:
#        print(f"  Brand: {brand}")
```

In [8]:

```
# Two rows with the same variant_id may have different size_label
rows2 = df[['variant_id', 'size_label']].groupby('variant_id').filter(lambda x: x['size_label'].nunique() != 1)
rows2
```

Out[8]:

	variant_id	size_label
3	14577988-17	one size
44	15626805-19	m
66	14639254-17	one size
96	14702959-17	one size
126	14011681-17	one size
...	...	...
3097577	17659877-17	40 cm
3097580	17482131-20	6 us
3097589	17349232-17	均码
3097590	17174861-17	1个月
3097602	16883841-17	einheitsgröße

330002 rows × 2 columns

In [9]:

```
df[df['variant_id'] == '17482131-20']
```

Out[9]:

	variant_id	product_id	size_label	product_name	brand	color	age_group	gender	size_type	product_type
150573	17482131-20	17482131	6 us	verziertes nelle kleid	alice+olivia	rot	adult	female	regular	kleidung > kleider > tageskleider
208845	17482131-20	17482131	4 us	nelle embellished dress	alice+olivia	red	adult	female	regular	clothing > dresses > day dresses
476785	17482131-20	17482131	6 us	nelle ドレス	alice+olivia	レッド	adult	female	regular	ウェア > ワンピース&ドレス > デイドレス
2074104	17482131-20	17482131	6 us	vestido nell con apliques	alice+olivia	rojo	adult	female	regular	ropa > vestidos > vestidos de día
2780116	17482131-20	17482131	6 us	nelle embellished dress	alice+olivia	red	adult	female	regular	clothing > dresses > day dresses
3055113	17482131-20	17482131	6 us	nelle 缀饰连衣裙	alice+olivia	红色	adult	female	regular	服装 > 连衣裙 > 日常连衣裙
3097580	17482131-20	17482131	6 us	vestido nelle com aplicações	alice+olivia	vermelho	adult	female	regular	roupas > vestidos > vestido casual

In [10]:

```
df[df['variant_id'] == '17659877-17']
```

Out[10]:

	variant_id	product_id	size_label	product_name	brand	color	age_group	gender	size_type	product_type
337840	17659877-17	17659877	40厘米	first class 刺绣套头帽	alviero martini kids	中性色	kids	unisex	regular	男婴配饰 > 男婴帽子 > 男婴针织帽
618246	17659877-17	17659877	40 cm	gorro first class bordado	alviero martini kids	neutro	kids	unisex	regular	accesorios para bebé niño > sombreros y gorros...
1100272	17659877-17	17659877	40 cm	first class ビーニー	alviero martini kids	ニュートラル	kids	unisex	regular	ベビーボーイズ ファッション小物 > ベビーボーイズ 帽子 > ベビーボーイズ ニット帽
1205141	17659877-17	17659877	40cm	first class embroidered beanie	alviero martini kids	nude	kids	unisex	regular	accessoires für baby boys > mützen für baby bo...
2277245	17659877-17	17659877	40 cm	first class embroidered beanie	alviero martini kids	neutrals	kids	unisex	regular	baby boy accessories > baby hats > baby knitte...
3097577	17659877-17	17659877	40 cm	gorro first class com bordado	alviero martini kids	neutro	kids	unisex	regular	acessórios para bebê > chapéu para bebê\n > to...

In [11]:

```
df[df['variant_id'] == '12899929-20']
```

Out[11]:

	variant_id	product_id	size_label	product_name	brand	color	age_group	gender	size_type	product_type
99	12899929-20	12899929	30 waist	bandana slim jeans	424	blue	adult	male	regular	clothing > denim > slim-fit jeans
162452	12899929-20	12899929	30 waist	bandana slim jeans	424	blue	adult	unisex	regular	clothing > denim > slim-fit jeans
1399285	12899929-20	12899929	30 waist	bandana slim jeans	424 fairfax	blue	adult	male	regular	clothing > denim > slim-fit jeans
2917127	12899929-20	12899929	30 waist	bandana slim jeans	424	blue	adult	male	regular	clothing > denim > slim-fit jeans

In [12]:

```
# Two rows with the same variant_id may have different product_name
rows3 = df[['variant_id', 'product_name']].groupby('variant_id').filter(lambda x: x['product_name'].nunique() != 1)
rows3
```

```
Out[12]:
                variant_id                    product_name
      2        14516134-31    side embroidered birds sneakers
      4        14796330-22                   logo knitted top
      5        14924756-42     crystal embellished double ring
      6        14505130-20          embroidered detail jumper
      8        14838238-25        oversized lace-up ankle boots
    ...                ...                              ...
3097600        17405854-29            dazed cowboy ankle boots
3097601        16720244-23   jogginghose mit reißverschlüssen
3097602        16883841-17          klassischer schal mit logo
3097603        17289151-20              logo-patch sweatpants
3097604        17330145-20          flared faux leather trousers

2006870 rows × 2 columns
```

```
In [13]: df[df['variant_id'] == '14838238-25']
```

Out[13]:

	variant_id	product_id	size_label	product_name	brand	color	age_group	gender	size_type	product_type
8	14838238-25	14838238	40 it	oversized lace-up ankle boots	alexander mcqueen	black	adult	male	regular	shoes > boots
341709	14838238-25	14838238	40 it	массивные ботинки на шнуровке	alexander mcqueen	черный	adult	male	regular	обувь > сапоги
349193	14838238-25	14838238	40 it	レースアップ アンクルブーツ	alexander mcqueen	ブラック	adult	male	regular	シューズ > ブーツ
670335	14838238-25	14838238	40 it	厚底系带及踝靴	alexander mcqueen	黑色	adult	male	regular	鞋履 > 靴子
847633	14838238-25	14838238	40 it	오버사이즈 레이스 업 앵클 부츠	alexander mcqueen	블랙	adult	male	regular	슈즈 > 부츠
939659	14838238-25	14838238	40 it	stiefel mit schnürung	alexander mcqueen	schwarz	adult	male	regular	schuhe > stiefel
1177272	14838238-25	14838238	40 it	bottines oversize à lacets	alexander mcqueen	noir	adult	male	regular	chaussures > bottes
1966826	14838238-25	14838238	40 it	ankle boot oversized com cadarço	alexander mcqueen	preto	adult	male	regular	sapatos > botas
2009215	14838238-25	14838238	40 it	botines oversize con cordones	alexander mcqueen	negro	adult	male	regular	zapatos > botas
2693261	14838238-25	14838238	40 it	stivaletti stringati	alexander mcqueen	nero	adult	male	regular	scarpe > stivali

In [14]:

```
# Two rows with the same variant_id may have different color
rows4 = df[['variant_id', 'color']].groupby('variant_id').filter(lambda x: x['color'].nunique() != 1)
rows4
```

Out[14]:

	variant_id	color
2	14516134-31	white
4	14796330-22	blue
5	14924756-42	gold
6	14505130-20	neutrals
8	14838238-25	black
...	...	...
3097600	17405854-29	schwarz
3097601	16720244-23	blau
3097602	16883841-17	nude
3097603	17289151-20	grau
3097604	17330145-20	braun

1838953 rows × 2 columns

In [15]:

```
rows4[rows4['variant_id'] == '14924756-42']
```

Out[15]:

	variant_id	color
5	14924756-42	gold
44909	14924756-42	oro
727703	14924756-42	dorado
1453303	14924756-42	gold
1796375	14924756-42	골드 톤
2107843	14924756-42	gold
2349353	14924756-42	or
2616386	14924756-42	золотистый
2693774	14924756-42	金色

In [16]:

```
# Two rows with the same variant_id may have different age_group
rows5 = df[['variant_id', 'age_group']].groupby('variant_id').filter(lambda x: x['age_group'].nunique() != 1)
rows5
```

Out[16]:

	variant_id	age_group
967	12143829-27	kids
2538	12397486-19	kids
3860	12359073-31	kids
3879	12787476-23	kids
3881	12868163-28	kids
...	...	...
3090537	12143829-23	kids
3092673	12472969-27	kids
3094788	11974405-18	kids
3095301	13221865-19	adult
3096446	11790517-22	kids

7121 rows × 2 columns

```
In [17]: df[df['variant_id'] == '13221865-19']
```

Out[17]:

	variant_id	product_id	size_label	product_name	brand	color	age_group	gender	size_type	product_type
733868	13221865-19	13221865	39 eu	teen lace-up sneakers	2 star kids	white	kids	unisex	regular	teen girl shoes > teen trainers
2259924	13221865-19	13221865	39 eu	teen lace-up sneakers	2 star kids	white	adult	unisex	regular	teen girl shoes > teen trainers
3095301	13221865-19	13221865	39 eu	2 star kids 2sb1303 bianco/nero cotton	2 star kids	white	adult	unisex	regular	teen girl shoes > teen trainers

```
In [18]: df[df['variant_id'] == '12143829-27']
```

Out[18]:

	variant_id	product_id	size_label	product_name	brand	color	age_group	gender	size_type	product_type
967	12143829-27	12143829	12 yrs	classic fitted dress	andorine	black	kids	unisex	regular	girls clothing > dresses > casual dresses
41982	12143829-27	12143829	12 yrs	classic fitted dress	andorine	black	kids	unisex	regular	girls clothing > dresses > casual dresses
274054	12143829-27	12143829	12 ans	robe cintrée	andorine	noir	kids	unisex	regular	vêtements fille > robes fille > robes décontra...
304124	12143829-27	12143829	12 j.	kleid mit langen ärmeln	andorine	schwarz	kids	unisex	regular	kleidung für mädchen > kleider für mädchen > f...
342080	12143829-27	12143829	12 yrs	classic fitted dress	andorine	black	kids	unisex	regular	girls clothing > girls dresses > girls casual ...
592645	12143829-27	12143829	12岁	拉链连衣裙	andorine	黑色	kids	unisex	regular	女童服装 > 女童连衣裙 > 女童休闲连衣裙
1271296	12143829-27	12143829	12 лет	классическое платье свободного кроя	andorine	черный	kids	unisex	regular	одежда для девочек (2-12 лет) > платья для дев...
1295768	12143829-27	12143829	12 años	vestido ajustado clásico	andorine	negro	kids	unisex	regular	ropa para niña > vestidos para niña > vestidos...
1928569	12143829-27	12143829	12歳	カジュアルワンピース	andorine	ブラック	kids	unisex	regular	ガールズ ウェア > ガールズ ワンピース > ガールズ カジュアルワンピース
1944651	12143829-27	12143829	12 세	클래식 핏 드레스	andorine	블랙	kids	unisex	regular	여아 / 의류 > 여아 / 드레스 > 여아 / 캐주얼 드레스
2126881	12143829-27	12143829	12 yrs	classic fitted dress	andorine	black	kids	unisex	regular	girls clothing > girls dresses > girls casual ...
2827482	12143829-27	12143829	12 anni	vestito classico	andorine	nero	kids	unisex	regular	abbigliamento bambina > abiti bambina > abiti ...
3011448	12143829-27	12143829	12 yrs	classic fitted dress	andorine	black	adult	unisex	regular	girls clothing > girls dresses > girls casual ...

In [19]:

```
df[['variant_id', 'age_group', 'product_type']][df['variant_id'] == '12787476-23']
```

Out[19]:

	variant_id	age_group	product_type
3879	12787476-23	kids	boys shoes > boys trainers
2019990	12787476-23	adult	boys shoes > boys trainers
2906527	12787476-23	kids	boys shoes > boys trainers

In [20]:

```
# Two rows with the same variant_id may have different gender
rows6 = df[['variant_id', 'gender']].groupby('variant_id').filter(lambda x: x['gender'].nunique() != 1)
rows6
```

Out[20]:

	variant_id	gender
82	12133372-23	male
99	12899929-20	male
102	12330066-19	female
216	12969981-17	female
220	15520017-17	female
...	...	...
3097236	13311763-22	male
3097237	12822675-29	unisex
3097344	12620014-17	female
3097581	11415056-21	male
3097594	13178707-21	female

227354 rows × 2 columns

In [21]:

```
df[['variant_id', 'gender', 'product_name', 'product_type']][df['variant_id'] == '12899929-20']
```

Out[21]:

	variant_id	gender	product_name	product_type
99	12899929-20	male	bandana slim jeans	clothing > denim > slim-fit jeans
162452	12899929-20	unisex	bandana slim jeans	clothing > denim > slim-fit jeans
1399285	12899929-20	male	bandana slim jeans	clothing > denim > slim-fit jeans
2917127	12899929-20	male	bandana slim jeans	clothing > denim > slim-fit jeans

In [22]:

```
df[['variant_id', 'gender', 'product_name', 'product_type']][df['variant_id'] == '12822675-29']
```

Out[22]:

	variant_id	gender	product_name	product_type
735932	12822675-29	unisex	424 fairfax x brandblack runners sneakers	shoes > sneakers > low-tops
1792886	12822675-29	unisex	424 fairfax x brandblack runners sneakers	shoes > sneakers > hi-tops
1934996	12822675-29	male	424 fairfax x brandblack runners sneakers	shoes > sneakers > low-tops
2921505	12822675-29	male	424 fairfax x brandblack runners sneakers	shoes > sneakers > low-tops
3097237	12822675-29	unisex	424 fairfax x brandblack runners sneakers	shoes > sneakers > hi-tops

In [23]:

```
df[['variant_id', 'gender', 'product_name', 'product_type']][df['variant_id'] == '12620014-17']
```

Out[23]:

	variant_id	gender	product_name	product_type
213212	12620014-17	female	'hungry snake' ohrringe aus sterlingsilber	schmuck > ohrringe
660788	12620014-17	female	hungry snake polished sterling silver earrings	jewelry > earrings
707283	12620014-17	female	pendientes hungry snake	joyas > pendientes
765697	12620014-17	female	饥饿蛇形耳环	珠宝 > 耳环
821233	12620014-17	female	hungry snake polished sterling silver earrings	jewellery > earrings
843766	12620014-17	female	silver hungry snake earrings	jewelry > fine earrings
972886	12620014-17	female	hungry snake sterling silver hoop earrings	fine jewellery > fine earrings
1051274	12620014-17	female	orecchini in argento sterling hungry snake	gioielli > orecchini
1129901	12620014-17	female	par de brincos hungry snake de prata	bijoux > brincos
1197684	12620014-17	female	sterling silver hungry snake earrings	fine jewellery > fine earrings
1241478	12620014-17	female	hungry snake sterling silver hoop earrings	jewellery > earrings
1272828	12620014-17	female	sterling silver hungry snake earrings	jewellery > earrings
1791877	12620014-17	female	boucles d'oreilles hungry snake	bijoux > boucles d'oreilles
1988493	12620014-17	unisex	sterling silver hungry snake earrings	fine jewellery > fine earrings
2154251	12620014-17	female	hungry snake polished sterling silver earrings	jewellery > earrings
2538383	12620014-17	female	헝그리 스네이크 스털링 실버 이어링	주얼리 > 이어링
2806439	12620014-17	female	sterling silver hungry snake earrings	fine jewellery > fine earrings
2842071	12620014-17	female	sterling silver hungry snake earrings	jewelry > fine earrings
2852171	12620014-17	female	hungry snake polished sterling silver earrings	jewellery > earrings
3046804	12620014-17	female	silver hungry snake ピアス	ジュエリー > ピアス＆イヤリング
3097344	12620014-17	female	серьги 'hungry snake'	украшения > серьги

In [24]:

```
# Two rows with the same variant_id may have different size_type
```

```
rows7 = df[['variant_id', 'size_type']].groupby('variant_id').filter(lambda x: x['size_type'].nunique() != 1)
rows7
```

Out[24]:

variant_id	size_type

In [25]:

```
# size_type is the same in the whole database
df['size_type'].unique()
```

Out[25]:

```
array(['regular'], dtype=object)
```

In [26]:

```
# Two rows with the same variant_id may have different product_type
rows8 = df[['variant_id', 'product_type']].groupby('variant_id').filter(lambda x: x['product_type'].nunique() != 1)
rows8
```

Out[26]:

	variant_id	product_type
2	14516134-31	shoes > trainers
4	14796330-22	clothing > knitwear > knitted tops
5	14924756-42	jewellery > rings
6	14505130-20	clothing > sweatshirts & knitwear > jumpers
8	14838238-25	shoes > boots
...	...	...
3097600	17405854-29	schuhe > stiefel & stiefeletten
3097601	16720244-23	kleidung > hosen > jogginghosen
3097602	16883841-17	accessoires > schals & halstücher
3097603	17289151-20	kleidung > hosen > jogginghosen
3097604	17330145-20	kleidung > hosen > schlaghosen

2025837 rows × 2 columns

In [27]:

```
df[['variant_id', 'product_name', 'product_type']][df['variant_id'] == '14516134-31']
```

Out[27]:

	variant_id	product_name	product_type
2	14516134-31	side embroidered birds sneakers	shoes > trainers
277431	14516134-31	sneakers con ricamo	scarpe > sneakers
692981	14516134-31	кроссовки с вышивкой	обувь > кроссовки
937329	14516134-31	사이드 자수 스니커즈	슈즈 > 스니커즈
960288	14516134-31	baskets à détails brodés	chaussures > baskets
1749308	14516134-31	zapatillas con detalles de pájaros	zapatos > zapatillas

In [28]:

```
df[['variant_id', 'product_name', 'product_type']][df['variant_id'] == '12620014-17']
```

Out[28]:

	variant_id	product_name	product_type
213212	12620014-17	'hungry snake' ohrringe aus sterlingsilber	schmuck > ohrringe
660788	12620014-17	hungry snake polished sterling silver earrings	jewelry > earrings
707283	12620014-17	pendientes hungry snake	joyas > pendientes
765697	12620014-17	饥饿蛇形耳环	珠宝 > 耳环
821233	12620014-17	hungry snake polished sterling silver earrings	jewellery > earrings
843766	12620014-17	silver hungry snake earrings	jewelry > fine earrings
972886	12620014-17	hungry snake sterling silver hoop earrings	fine jewellery > fine earrings
1051274	12620014-17	orecchini in argento sterling hungry snake	gioielli > orecchini
1129901	12620014-17	par de brincos hungry snake de prata	bijoux > brincos
1197684	12620014-17	sterling silver hungry snake earrings	fine jewellery > fine earrings
1241478	12620014-17	hungry snake sterling silver hoop earrings	jewellery > earrings
1272828	12620014-17	sterling silver hungry snake earrings	jewellery > earrings
1791877	12620014-17	boucles d'oreilles hungry snake	bijoux > boucles d'oreilles
1988493	12620014-17	sterling silver hungry snake earrings	fine jewellery > fine earrings
2154251	12620014-17	hungry snake polished sterling silver earrings	jewellery > earrings
2538383	12620014-17	헝그리 스네이크 스털링 실버 이어링	주얼리 > 이어링
2806439	12620014-17	sterling silver hungry snake earrings	fine jewellery > fine earrings
2842071	12620014-17	sterling silver hungry snake earrings	jewelry > fine earrings
2852171	12620014-17	hungry snake polished sterling silver earrings	jewellery > earrings
3046804	12620014-17	silver hungry snake ピアス	ジュエリー > ピアス＆イヤリング
3097344	12620014-17	серьги 'hungry snake'	украшения > серьги

In [ ]: