How Important is Data Quality for Data Engineer?

(づ｡◕‿‿◕｡)づ

@manelbutterfly

Who Am I ?

- Data Engineer at Secret Sauce Partners Inc.
- Mozilla Representative
- Enthusiast about new Data technologies

"The only way to do great work is to love the work you do"
Steve Jobs

@manelbutterfly

What We Do?

secret sauce

PARTNERS

Transforming apparel & footwear shopping through data.

DXL MENS APPAREL

BARNEYS NEW YORK

Saks Fifth Avenue OFF 5TH

Google shopping

@manelbutterfly

NORDSTROM rack | HAUTELOOK

What We Do?

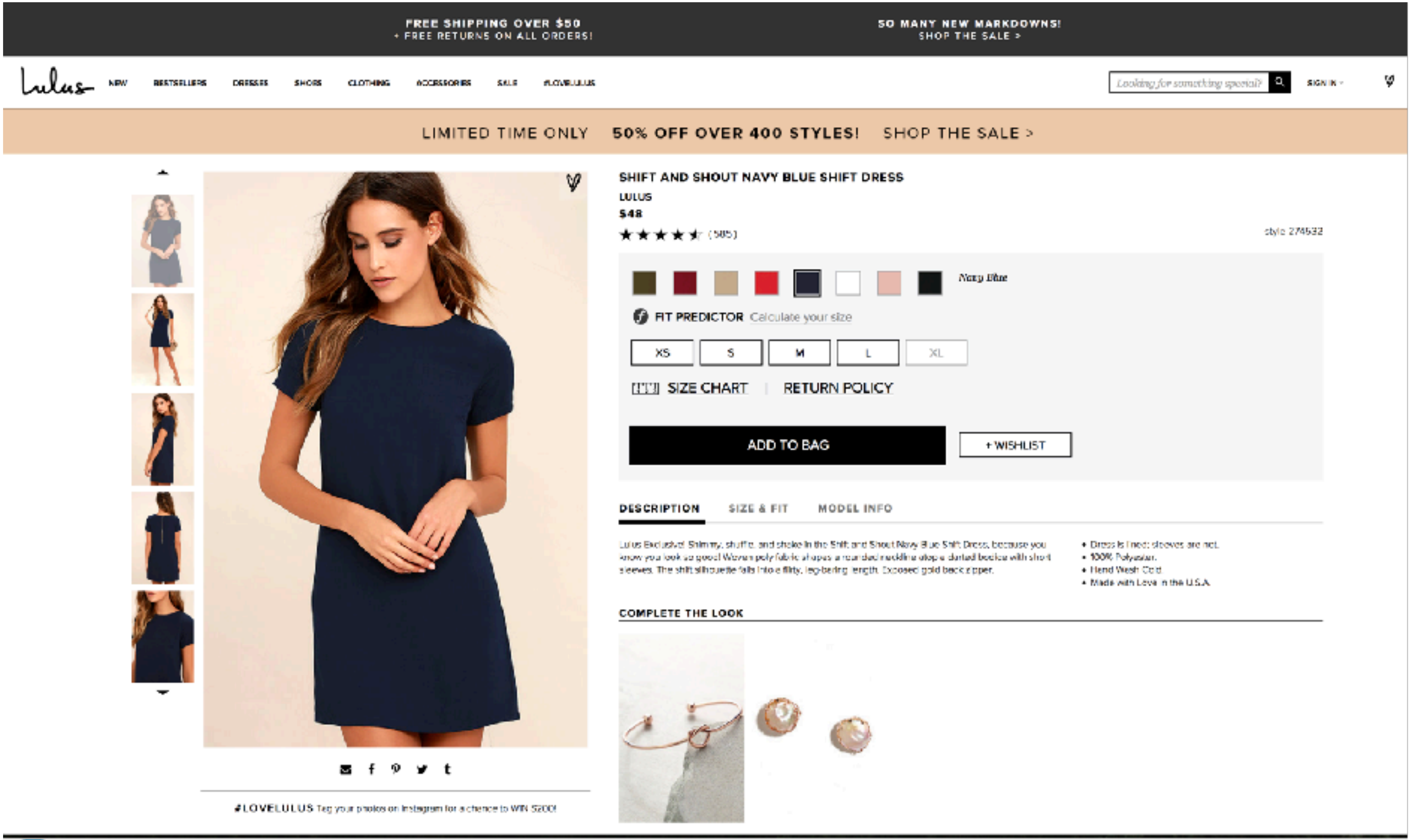
secret sauce

PARTNERS

Company's service

FIT PREDICTOR

What We Do?



FREE SHIPPING OVER $50
+ FREE RETURNS ON ALL ORDERS!

SO MANY NEW MARKDOWNS!
SHOP THE SALE >

LIMITED TIME ONLY **50% OFF OVER 400 STYLES!** SHOP THE SALE >

SHIFT AND SHOUT NAVY BLUE SHIFT DRESS
LULUS
$48
★★★★★ (505)
style 274532

Navy Blue

FIT PREDICTOR Calculate your size

XS S M L XL

SIZE CHART | RETURN POLICY

ADD TO BAG
+ WISHLIST

DESCRIPTION SIZE & FIT MODEL INFO

Lulus Exclusive! Shimmy, shuffle, and shake in the Shift and Shout Navy Blue Shift Dress, because you know you look so good! Woven poly fabric shapes a rounded neckline atop a darted bodice with short sleeves. The shift silhouette falls into a flirty, leg-baring length. Exposed gold back zipper.

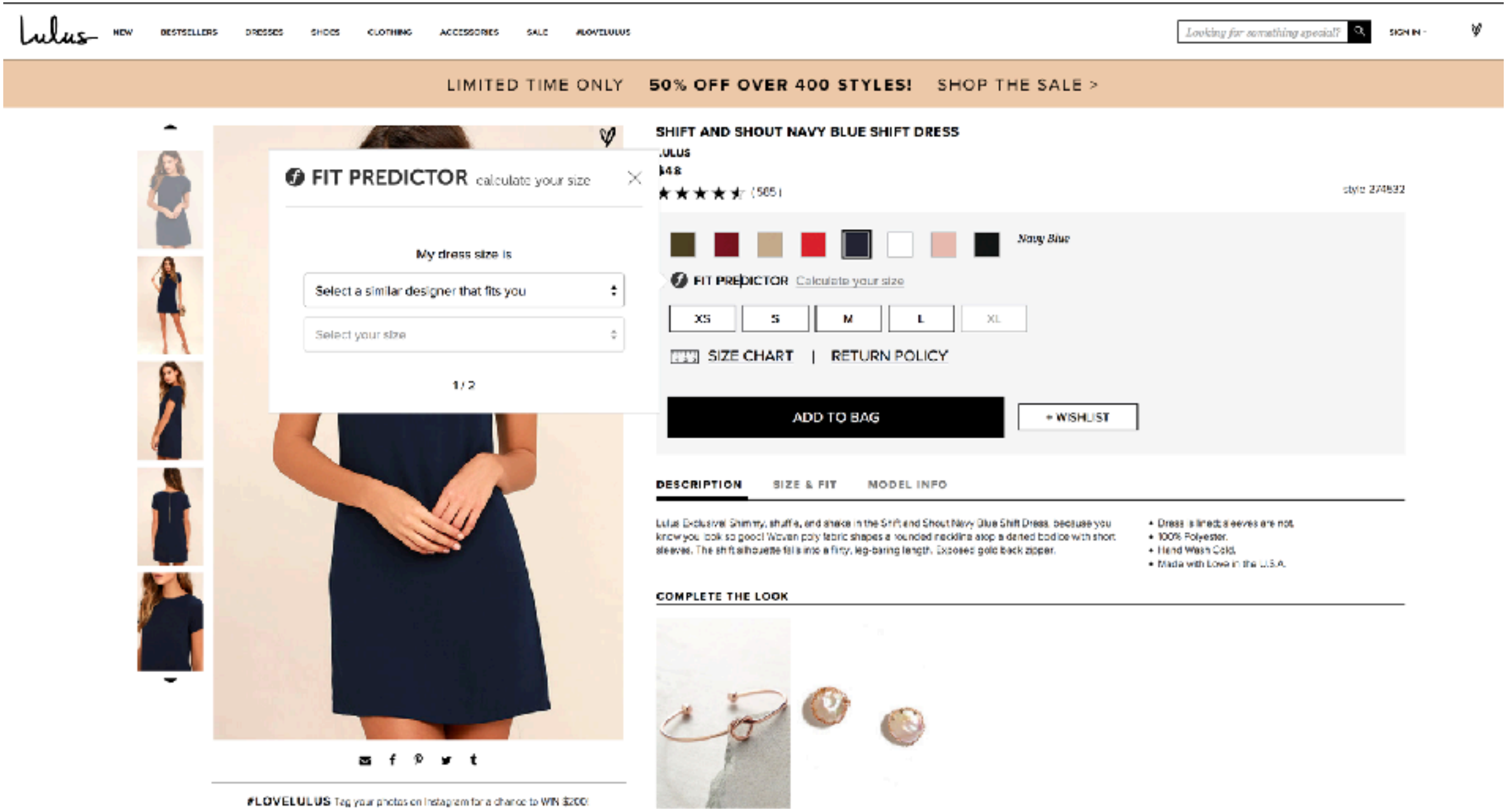
- Dress is lined; sleeves are not.
- 100% Polyester.
- Hand Wash Cold.
- Made with Love in the U.S.A.

COMPLETE THE LOOK

#LOVELULUS Tag your photos on Instagram for a chance to WIN $200!

@manelbutterfly

What We Do?



LIMITED TIME ONLY 50% OFF OVER 400 STYLES! SHOP THE SALE >

FIT PREDICTOR calculate your size

My dress size is

Select a similar designer that fits you

Select your size

1 / 2

SHIFT AND SHOUT NAVY BLUE SHIFT DRESS

LULUS

$48

★★★★½ (585)

style 274832

Navy Blue

FIT PREDICTOR Calculate your size

XS | S | M | L | XL

SIZE CHART | RETURN POLICY

ADD TO BAG

+ WISHLIST

DESCRIPTION SIZE & FIT MODEL INFO

Lulus Exclusive! Shimmy, shuffle, and shake in the Shift and Shout Navy Blue Shift Dress, because you know you look so good! Woven poly fabric shapes a rounded neckline atop a darted bodice with short sleeves. The shift silhouette falls into a flirty, leg-baring length. Exposed gold back zipper.

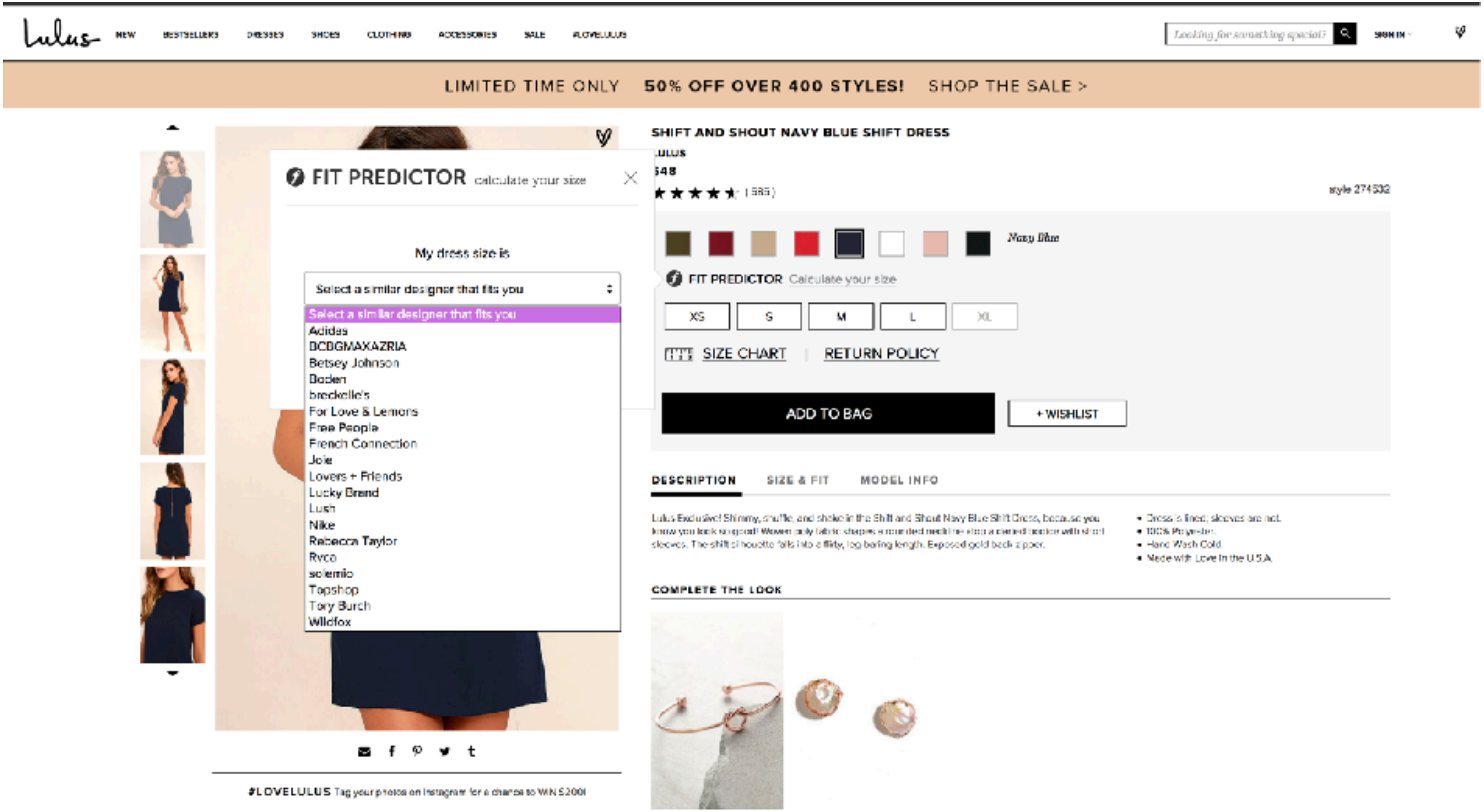
- Dress is lined; sleeves are not.
- 100% Polyester.
- Hand Wash Cold.
- Made with Love in the U.S.A.

COMPLETE THE LOOK

#LOVELULUS Tag your photos on Instagram for a chance to WIN $200!

@manelbutterfly

What We Do?



LIMITED TIME ONLY 50% OFF OVER 400 STYLES! SHOP THE SALE >

FIT PREDICTOR calculate your size

My dress size is

Select a similar designer that fits you

- Select a similar designer that fits you
- Adidas
- BCBGMAXAZRIA
- Betsey Johnson
- Boden
- brackelle's
- For Love & Lemons
- Free People
- French Connection
- Joie
- Lovers + Friends
- Lucky Brand
- Lush
- Nike
- Rebecca Taylor
- Rvca
- solemio
- Topshop
- Tory Burch
- Wildfox

SHIFT AND SHOUT NAVY BLUE SHIFT DRESS

LULUS

$48

★★★★★ (585)

style 274532

Navy Blue

FIT PREDICTOR Calculate your size

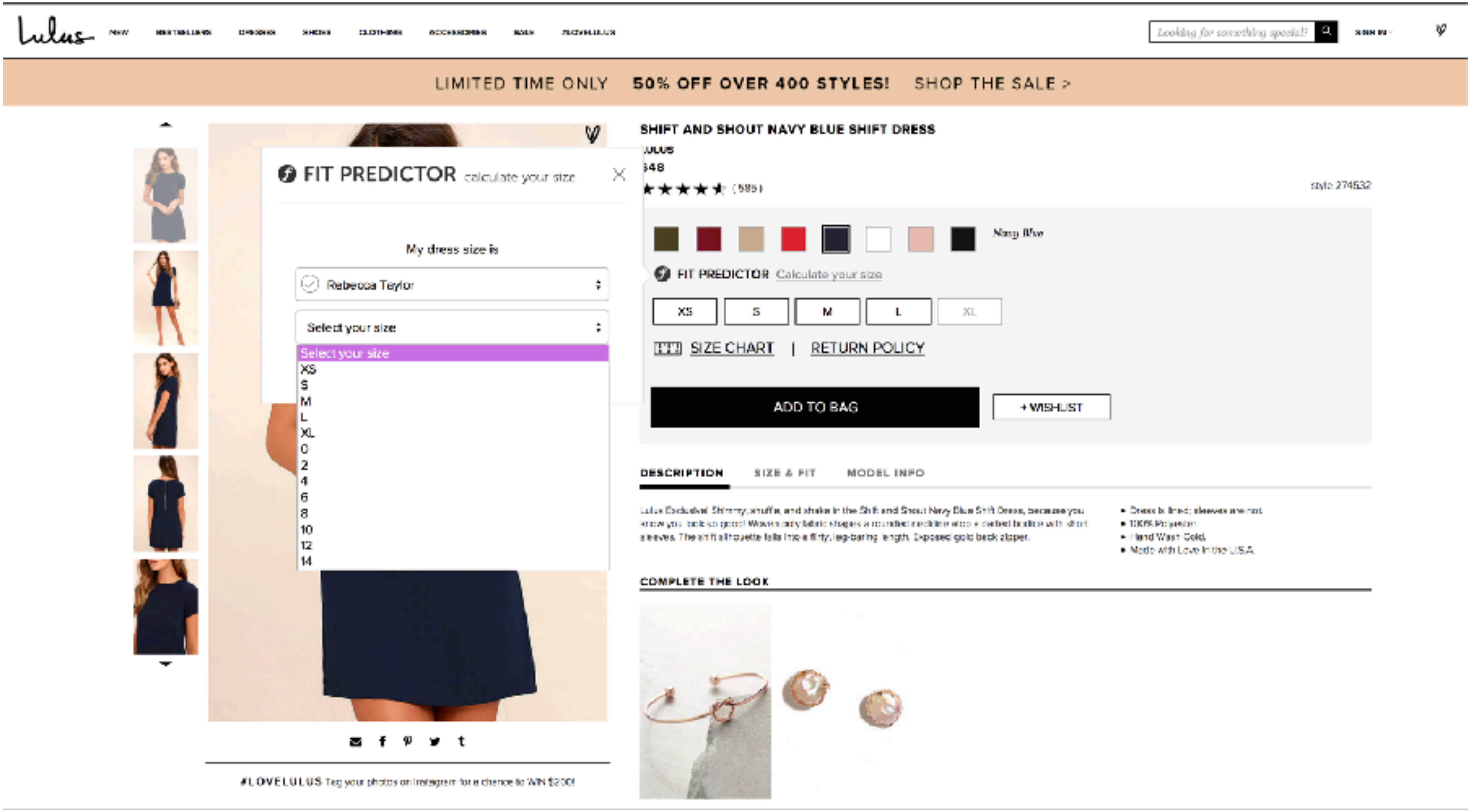
Lulus Exclusive! Shimmy, shuffle, and shake in the Shift and Shout Navy Blue Shift Dress, because you know you look so good! Woven poly fabric shapes a rounded neckline atop a darted bodice with short sleeves. The shift silhouette falls into a flirty, leg baring length. Exposed gold back zipper.

- Dress is lined; sleeves are not.
- 100% Polyester.
- Hand Wash Cold.
- Made with Love in the U.S.A.

COMPLETE THE LOOK

#LOVELULUS Tag your photos on Instagram for a chance to WIN $200!

What We Do?



Lulus NEW BESTSELLERS DRESSES SHOES CLOTHING ACCESSORIES SALE #LOVELULUS

Looking for something special?

SIGN IN

LIMITED TIME ONLY **50% OFF OVER 400 STYLES!** SHOP THE SALE >

FIT PREDICTOR calculate your size

My dress size is

Rebecca Taylor

Select your size

- Select your size
- XS
- S
- M
- L
- XL
- 0
- 2
- 4
- 6
- 8
- 10
- 12
- 14

SHIFT AND SHOUT NAVY BLUE SHIFT DRESS

LULUS

$48

★★★★★ (585)

style 274532

Navy Blue

FIT PREDICTOR Calculate your size

XS S M L XL

SIZE CHART | RETURN POLICY

ADD TO BAG

+ WISHLIST

DESCRIPTION SIZE & FIT MODEL INFO

Lulus Exclusive! Shimmy, shuffle, and shake in the Shift and Shout Navy Blue Shift Dress, because you know you look so good! Woven poly fabric shapes a rounded neckline atop a darted bodice with short sleeves. The shift silhouette falls into a flirty, leg-baring length. Exposed gold back zipper.

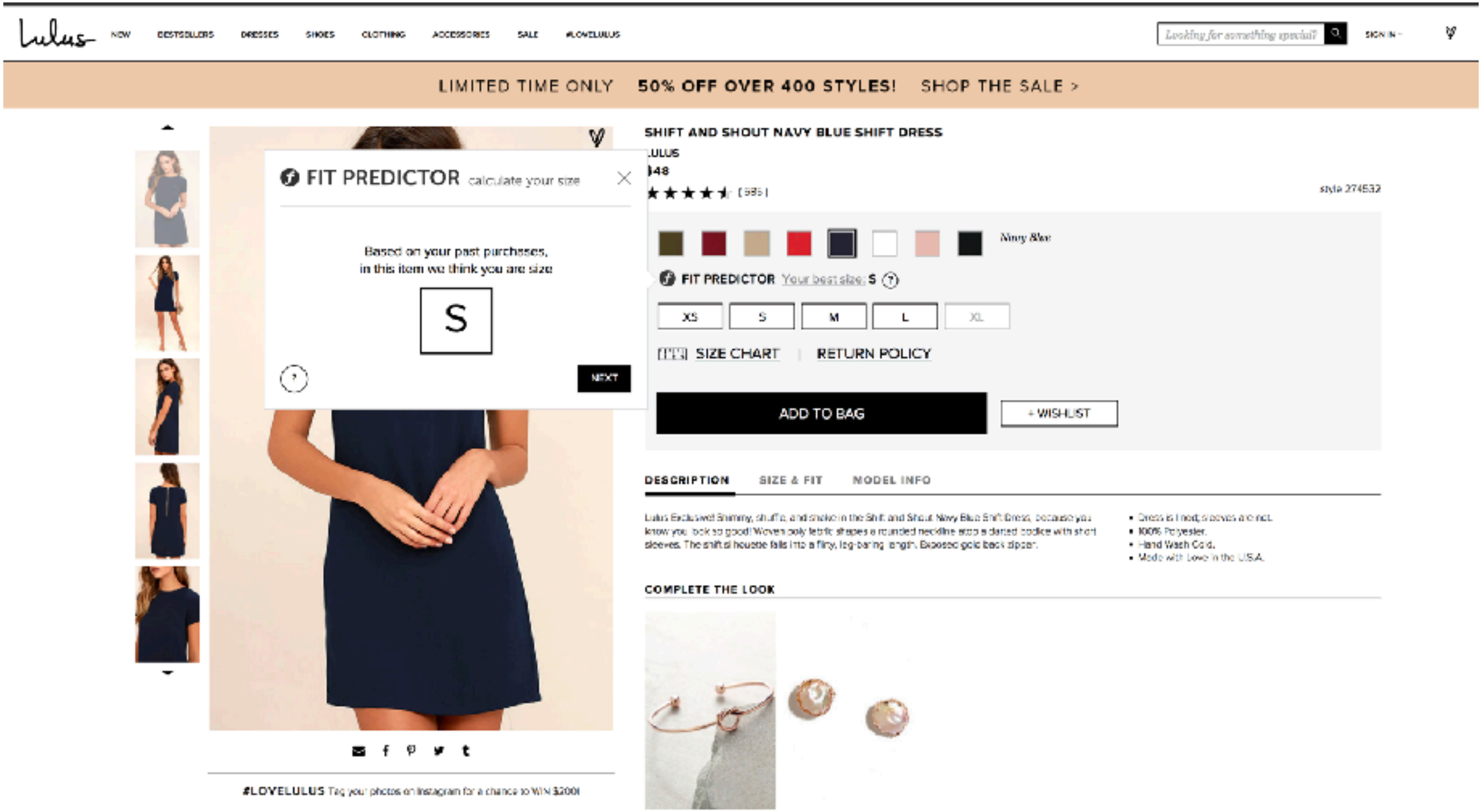
- Dress is lined; sleeves are not.
- 100% Polyester.
- Hand Wash Cold.
- Made with Love in the U.S.A.

COMPLETE THE LOOK

#LOVELULUS Tag your photos on Instagram for a chance to WIN $200!

@manelbutterfly

What We Do?



LIMITED TIME ONLY **50% OFF OVER 400 STYLES!** SHOP THE SALE >

FIT PREDICTOR calculate your size

Based on your past purchases,
in this item we think you are size

S

SHIFT AND SHOUT NAVY BLUE SHIFT DRESS

LULUS

$48

★★★★½ (595)

style 274532

Navy Blue

FIT PREDICTOR Your best size: **S**

XS S M L XL

SIZE CHART | RETURN POLICY

DESCRIPTION SIZE & FIT MODEL INFO

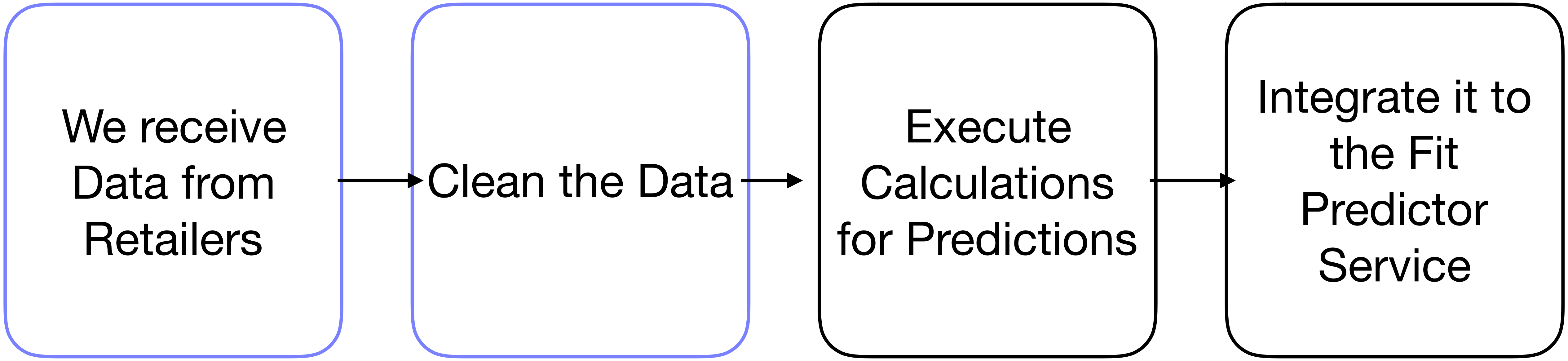
Lulus Exclusive! Shimmy, shuffle, and shake in the Shift and Shout Navy Blue Shift Dress, because you know you look so good! Woven poly fabric shapes a rounded neckline atop a darted bodice with short sleeves. The shift silhouette falls into a flirty, leg-baring length. Exposed gold back zipper.

- Dress is lined, sleeves are not.
- 100% Polyester.
- Hand Wash Cold.
- Made with Love in the U.S.A.

COMPLETE THE LOOK

#LOVELULUS Tag your photos on Instagram for a chance to WIN $200!

What We Do?



We receive Data from Retailers

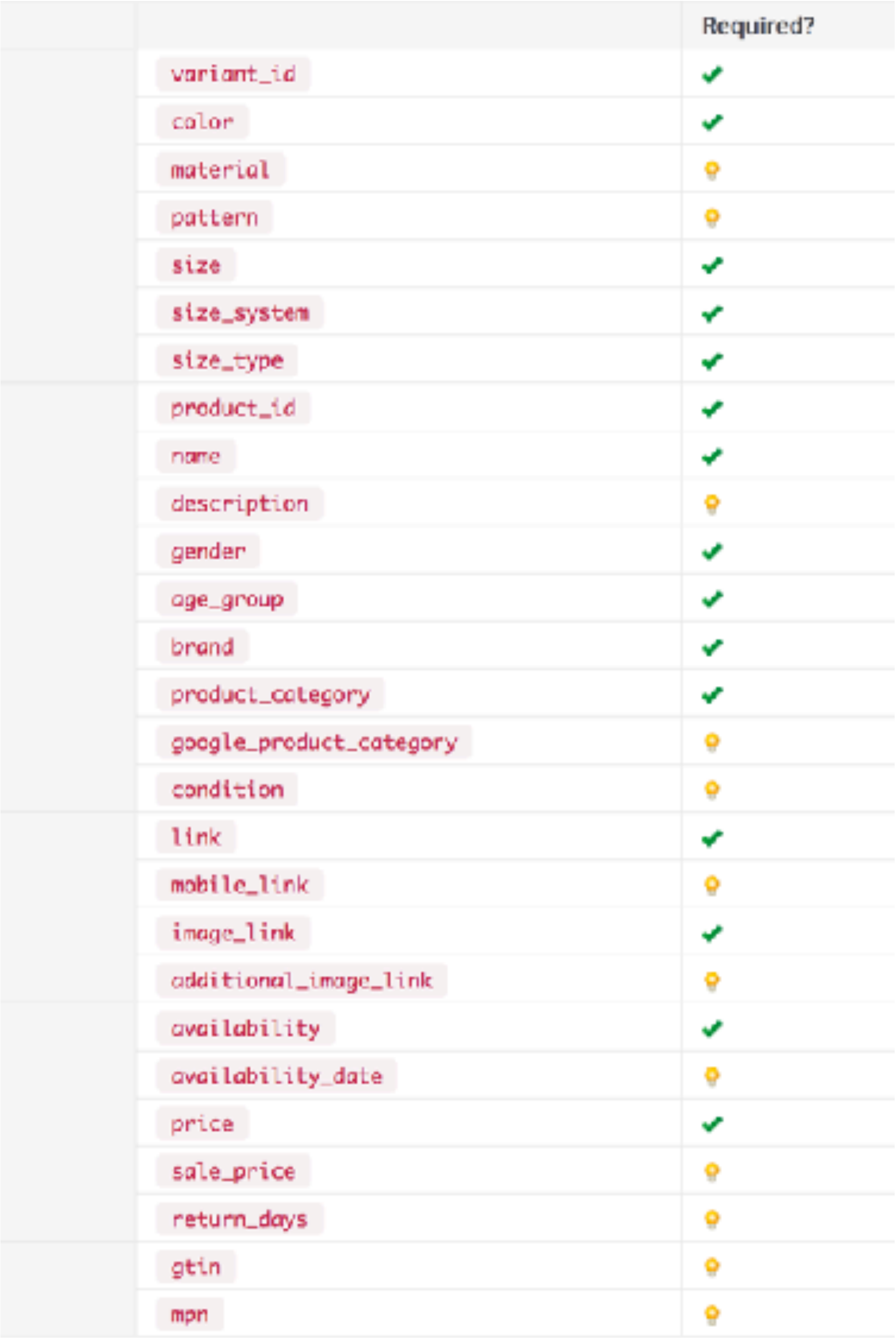
Clean the Data

Execute Calculations for Predictions

Integrate it to the Fit Predictor Service

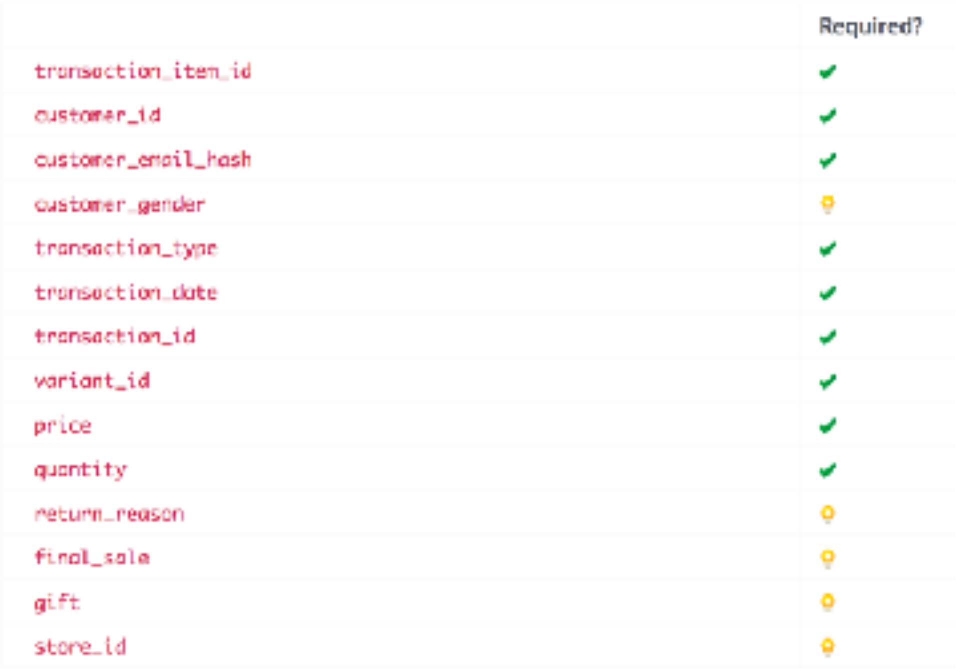
What We Do?

Product's Specification Requirements



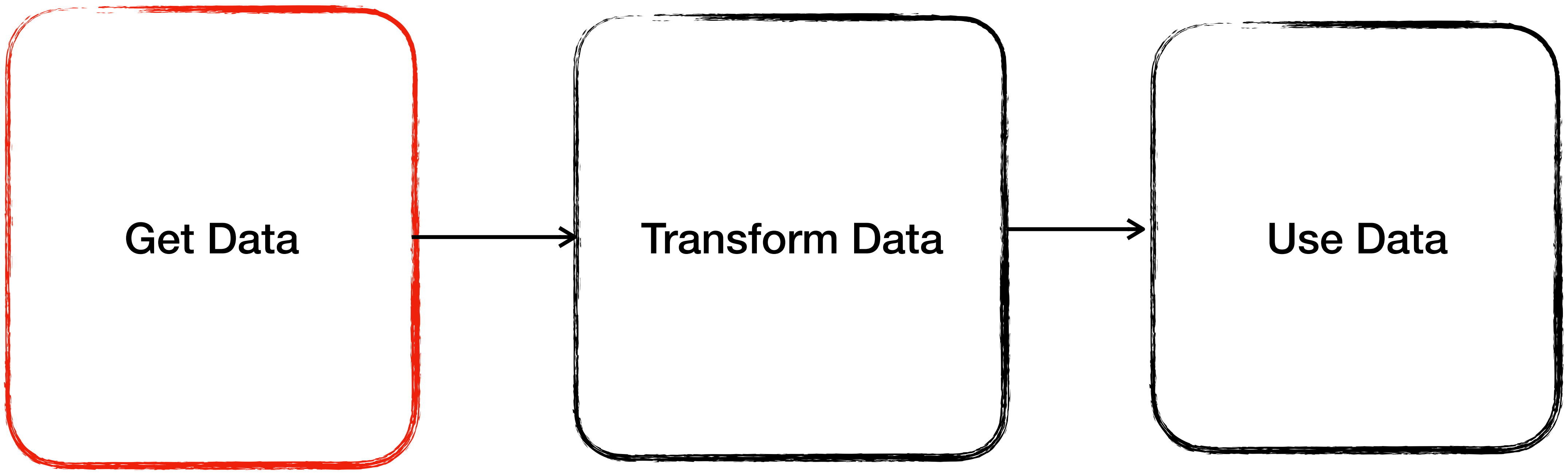
	Required?
variant_id	✓
color	✓
material	💡
pattern	💡
size	✓
size_system	✓
size_type	✓
product_id	✓
name	✓
description	💡
gender	✓
age_group	✓
brand	✓
product_category	✓
google_product_category	💡
condition	💡
link	✓
mobile_link	💡
image_link	✓
additional_image_link	💡
availability	✓
availability_date	💡
price	✓
sale_price	💡
return_days	💡
gtin	💡
mpn	💡

Transaction's Specification Requirements



	Required?
transaction_item_id	✓
customer_id	✓
customer_email_hash	✓
customer_gender	💡
transaction_type	✓
transaction_date	✓
transaction_id	✓
variant_id	✓
price	✓
quantity	✓
return_reason	💡
final_sale	💡
gift	💡
store_id	💡

Where is Data Quality?



@manelbutterfly

Where is Data Quality?

@manelbutterfly

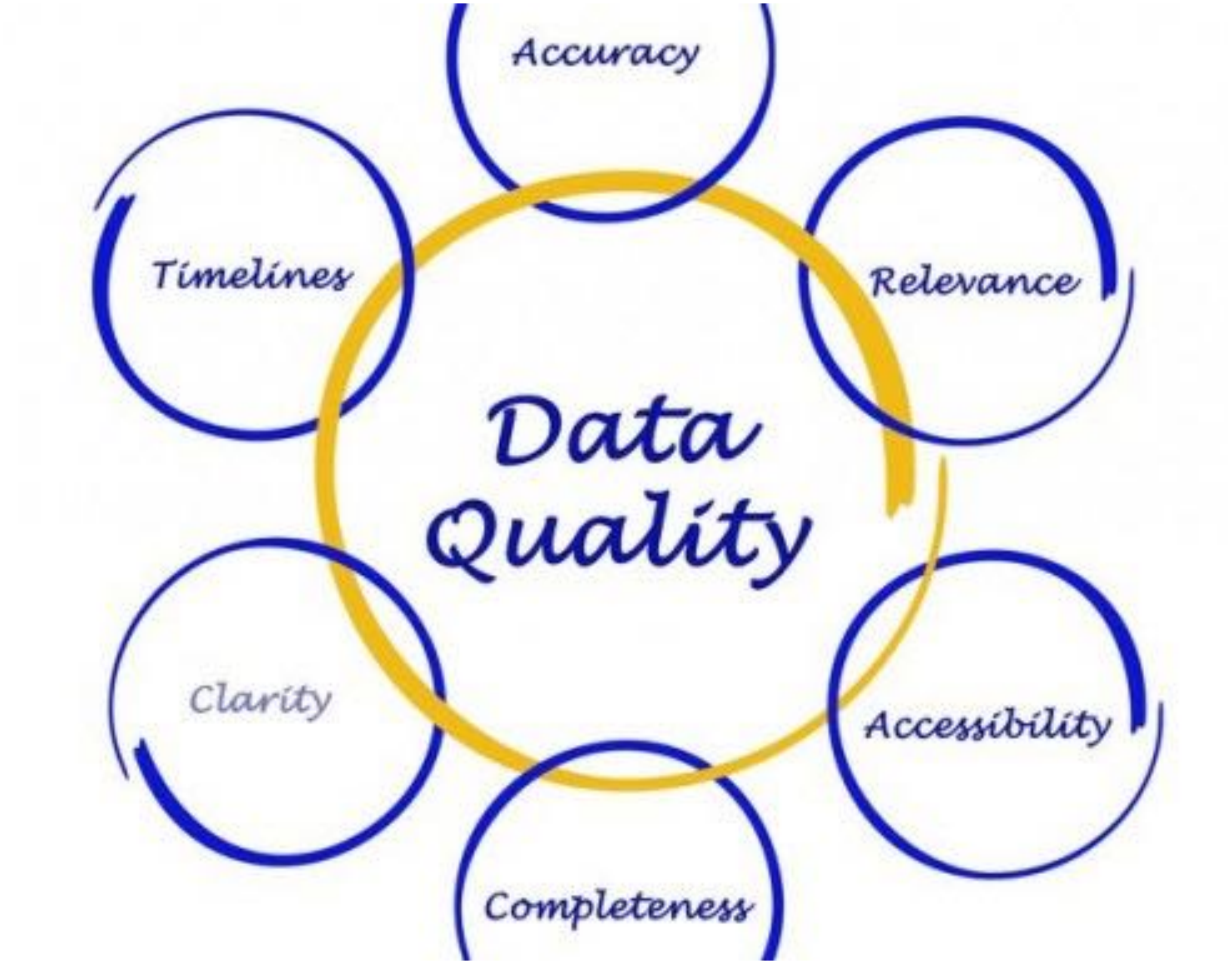
First Impression

═̿̿'̿'\̵͇̿̿\=(•̪●)=/̵͇̿̿/'̿̿═══

"Data don't make any sense,
we will have to resort to statistics."

@manelbutterfly

How Should I qualify this Data? ⊂•⊃_⊂•⊃



@manelbutterfly

But How I can ensure all of this?

¯_(ツ)_/¯

Profiling ✌.ʕʘ‿ʘʔ.✌

- ✦ Anomalies ?
- ✦ Content? Structure? Relation?
- ✦ Can this apply to our Business Logic?
- ✦ Missing values ?
- ✦ Distribution of required information

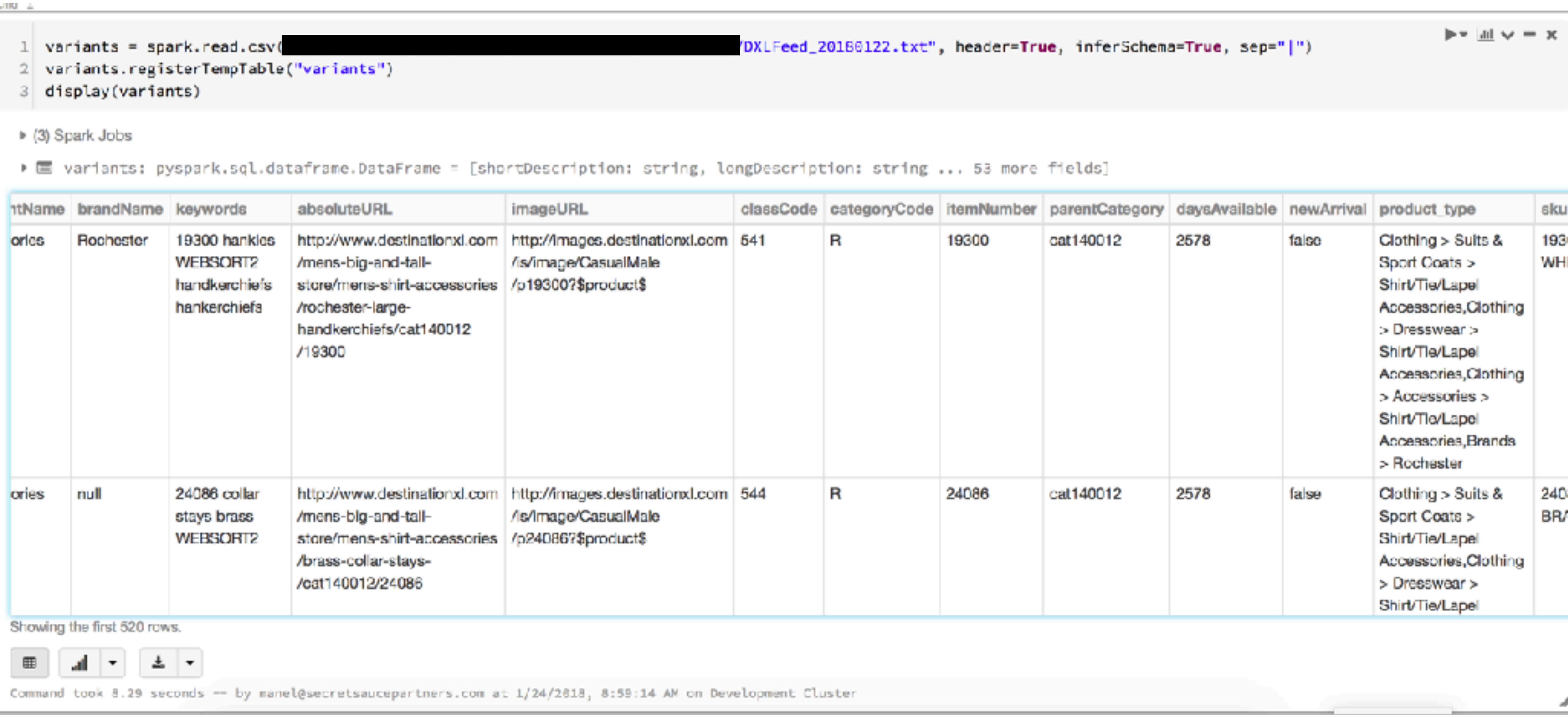
Profiling ✌.ʕʘ‿ʘʔ.✌

Spark profiling module

Simple query

Simple Statistics

Profiling ✌.ʕʘ‿ʘʔ.✌



```
variants = spark.read.csv([redacted]/DXLFeed_20180122.txt", header=True, inferSchema=True, sep="|")
variants.registerTempTable("variants")
display(variants)
```

▸ (3) Spark Jobs

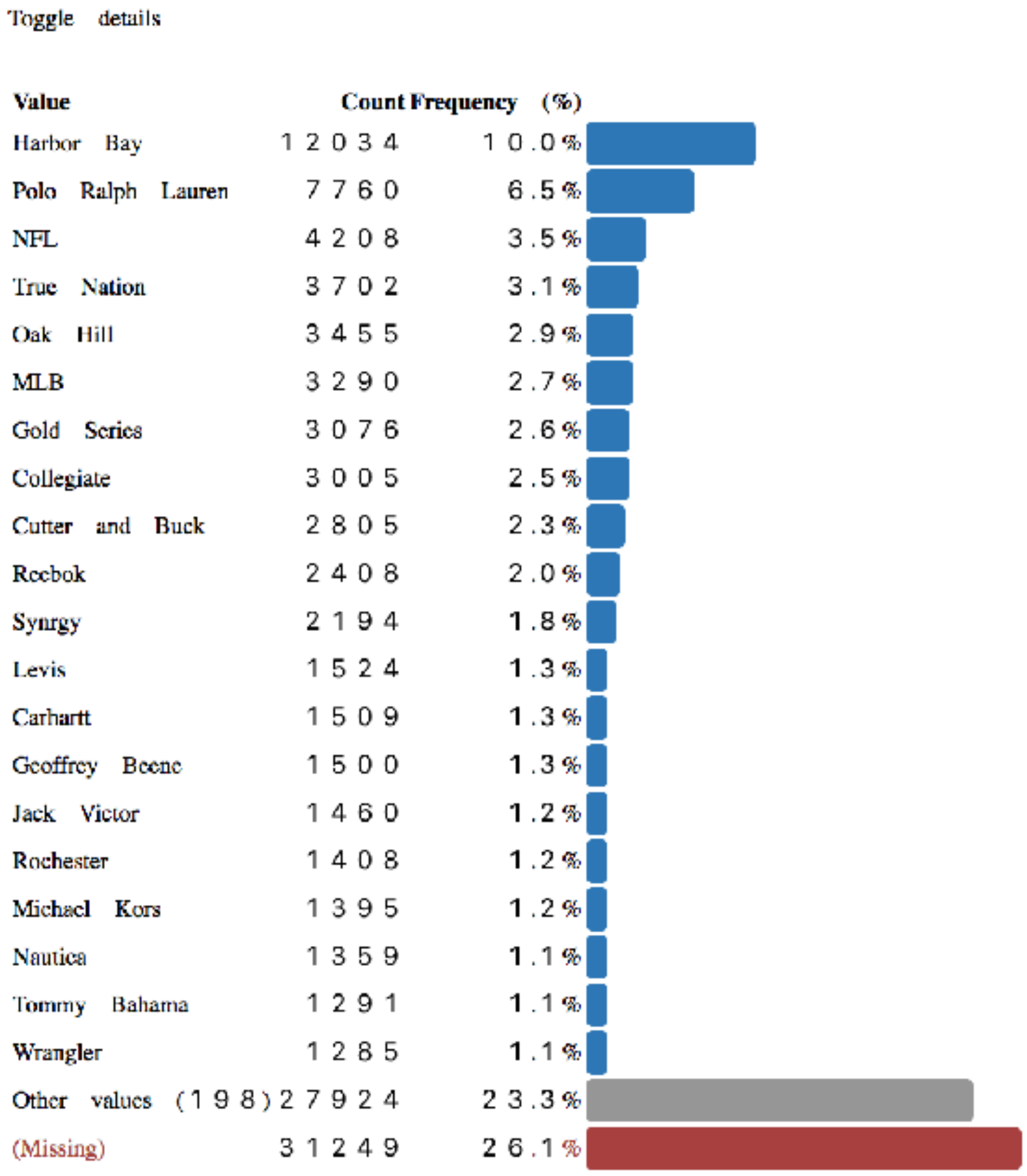
▸ variants: pyspark.sql.dataframe.DataFrame = [shortDescription: string, longDescription: string ... 53 more fields]

ntName	brandName	keywords	absoluteURL	imageURL	classCode	categoryCode	itemNumber	parentCategory	daysAvailable	newArrival	product_type	sku
orics	Rochester	19300 hankies WEBSORT2 handkerchiefs hankerchiefs	http://www.destinationxl.com/mens-big-and-tall-store/mens-shirt-accessories/rochester-large-handkerchiefs/cat140012/19300	http://images.destinationxl.com/is/image/CasualMale/p19300?$product$	541	R	19300	cat140012	2578	false	Clothing > Suits & Sport Coats > Shirt/Tie/Lapel Accessories,Clothing > Dresswear > Shirt/Tie/Lapel Accessories,Clothing > Accessories > Shirt/Tie/Lapel Accessories,Brands > Rochester	193 WH
ories	null	24086 collar stays brass WEBSORT2	http://www.destinationxl.com/mens-big-and-tall-store/mens-shirt-accessories/brass-collar-stays-/cat140012/24086	http://images.destinationxl.com/is/image/CasualMale/p24086?$product$	544	R	24086	cat140012	2578	false	Clothing > Suits & Sport Coats > Shirt/Tie/Lapel Accessories,Clothing > Dresswear > Shirt/Tie/Lapel	240 BR

Showing the first 520 rows.

```
Command took 8.29 seconds -- by manel@secretsaucepartners.com at 1/24/2018, 8:59:14 AM on Development Cluster
```

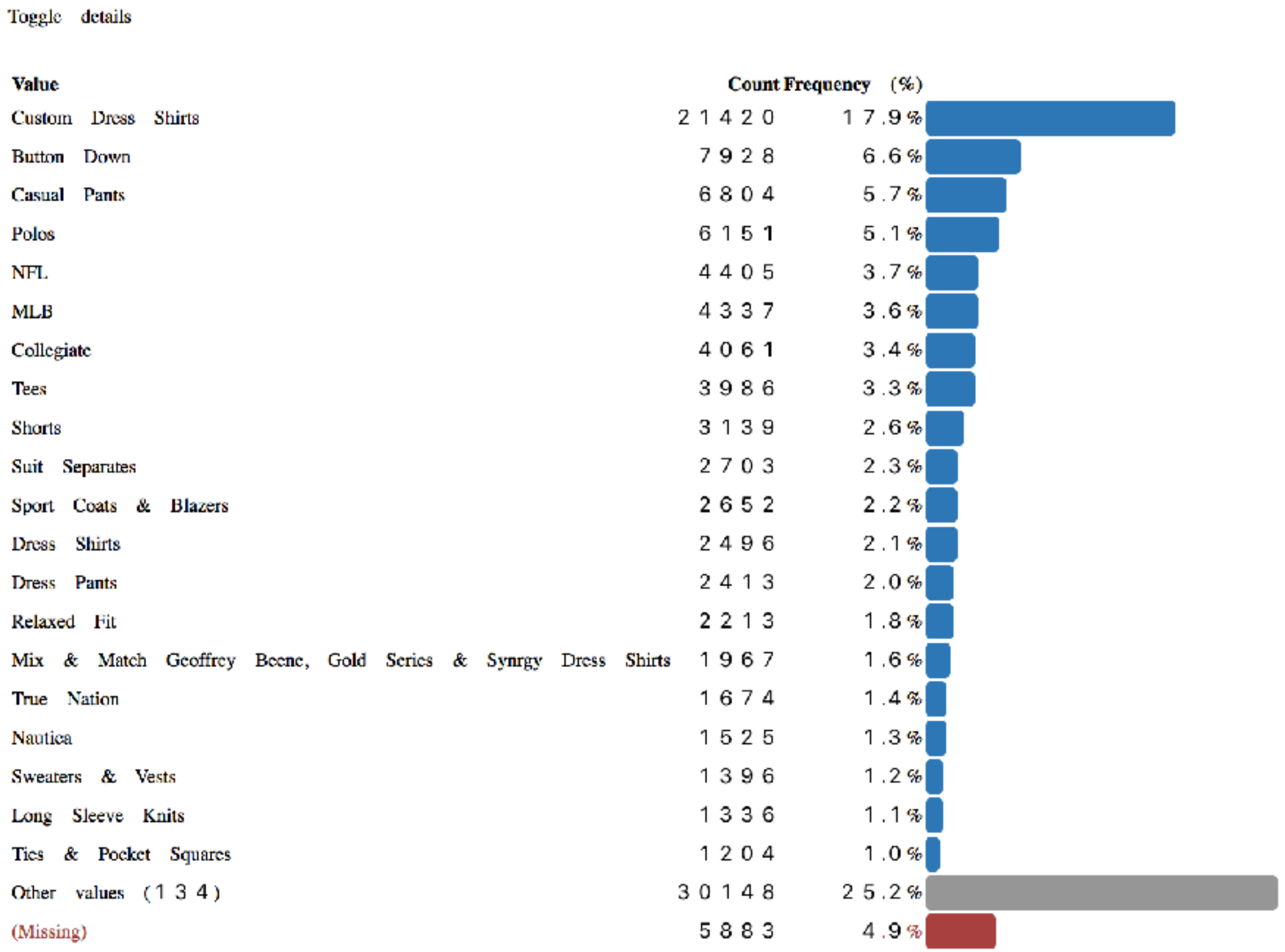
@manelbutterfly



Toggle details

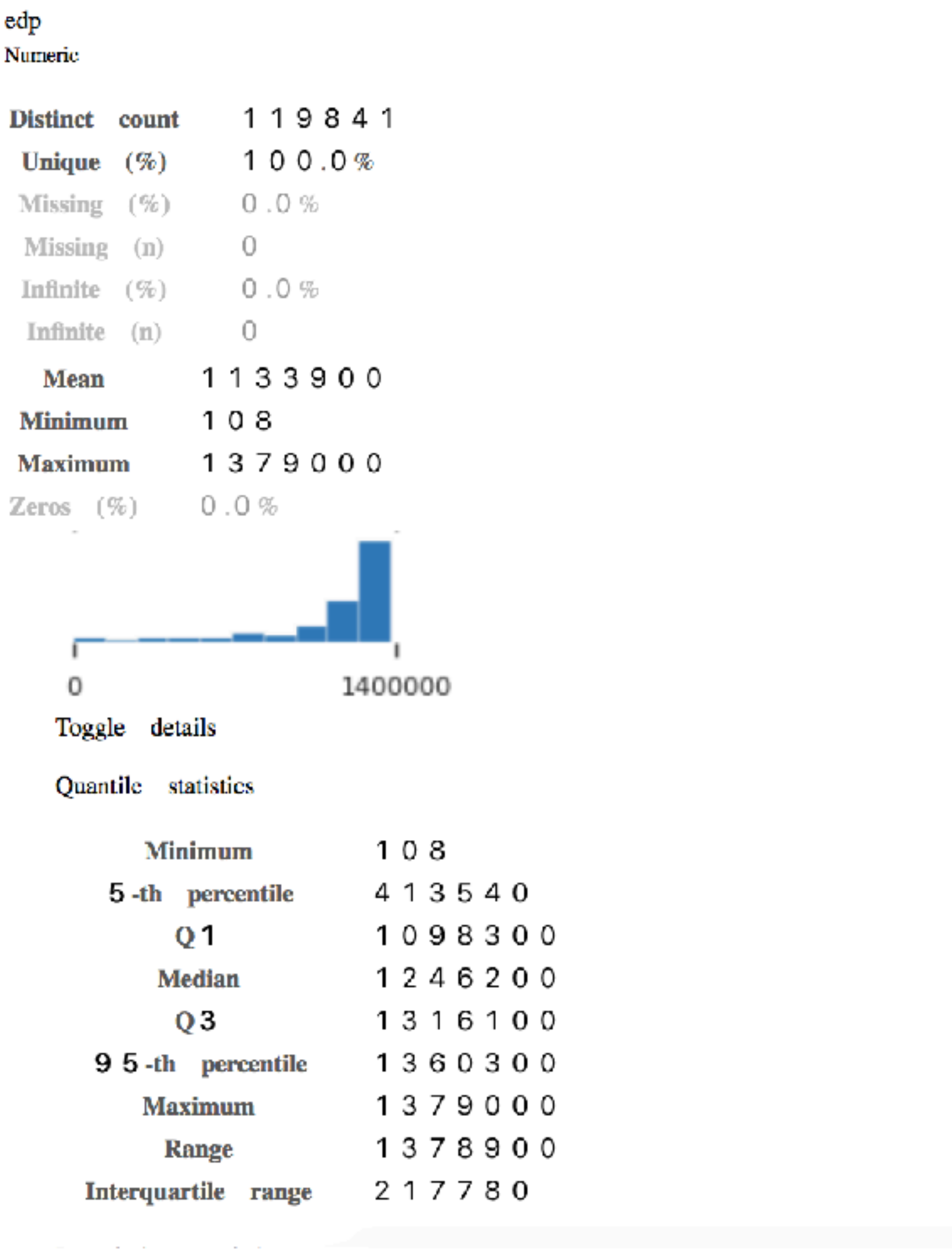
Value	Count	Frequency (%)
Harbor Bay	12034	10.0%
Polo Ralph Lauren	7760	6.5%
NFL	4208	3.5%
True Nation	3702	3.1%
Oak Hill	3455	2.9%
MLB	3290	2.7%
Gold Series	3076	2.6%
Collegiate	3005	2.5%
Cutter and Buck	2805	2.3%
Reebok	2408	2.0%
Synrgy	2194	1.8%
Levis	1524	1.3%
Carhartt	1509	1.3%
Geoffrey Beene	1500	1.3%
Jack Victor	1460	1.2%
Rochester	1408	1.2%
Michael Kors	1395	1.2%
Nautica	1359	1.1%
Tommy Bahama	1291	1.1%
Wrangler	1285	1.1%
Other values (198)	27924	23.3%
(Missing)	31249	26.1%

@manelbutterfly



Toggle details

Value	Count	Frequency (%)
Custom Dress Shirts	21420	17.9%
Button Down	7928	6.6%
Casual Pants	6804	5.7%
Polos	6151	5.1%
NFL	4405	3.7%
MLB	4337	3.6%
Collegiate	4061	3.4%
Tees	3986	3.3%
Shorts	3139	2.6%
Suit Separates	2703	2.3%
Sport Coats & Blazers	2652	2.2%
Dress Shirts	2496	2.1%
Dress Pants	2413	2.0%
Relaxed Fit	2213	1.8%
Mix & Match Geoffrey Beene, Gold Series & Synrgy Dress Shirts	1967	1.6%
True Nation	1674	1.4%
Nautica	1525	1.3%
Sweaters & Vests	1396	1.2%
Long Sleeve Knits	1336	1.1%
Ties & Pocket Squares	1204	1.0%
Other values (134)	30148	25.2%
(Missing)	5883	4.9%



edp

Numeric

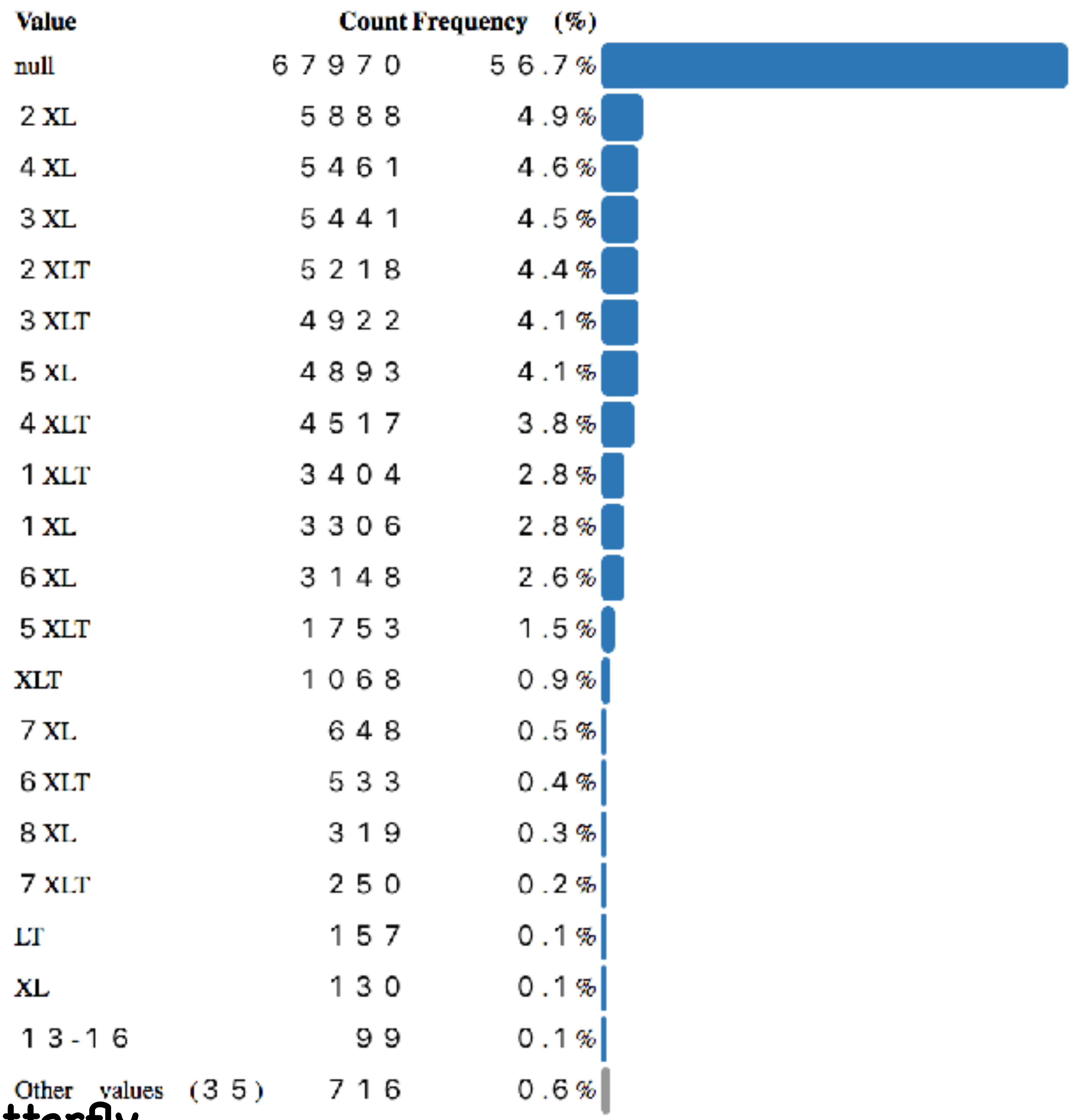
Distinct count	119841
Unique (%)	100.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Mean	1133900
Minimum	108
Maximum	1379000
Zeros (%)	0.0%

Toggle details

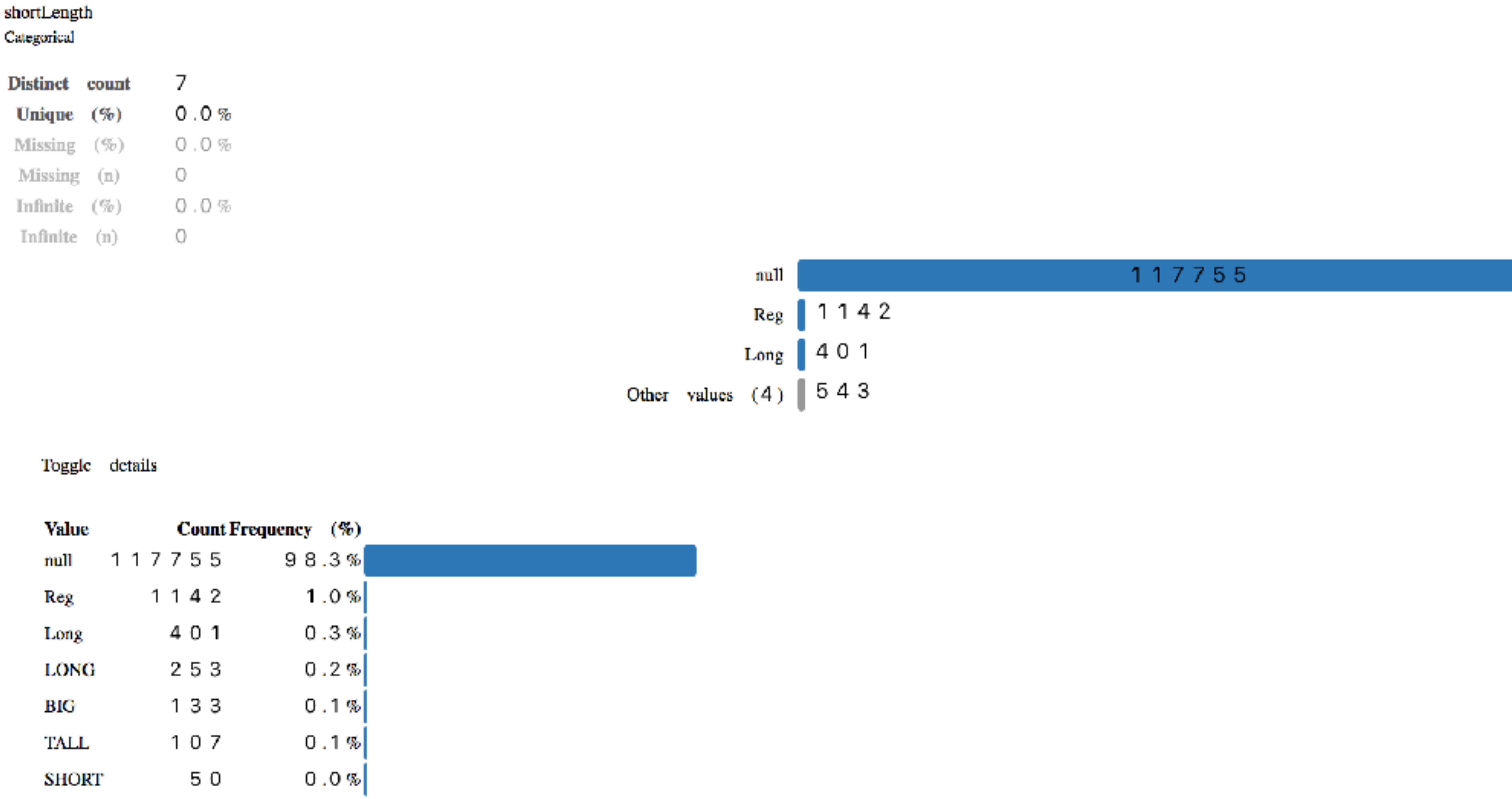
Quantile statistics

Minimum	108
5-th percentile	413540
Q1	1098300
Median	1246200
Q3	1316100
95-th percentile	1360300
Maximum	1379000
Range	1378900
Interquartile range	217780



Value	Count	Frequency (%)
null	67970	56.7%
2 XL	5888	4.9%
4 XL	5461	4.6%
3 XL	5441	4.5%
2 XLT	5218	4.4%
3 XLT	4922	4.1%
5 XL	4893	4.1%
4 XLT	4517	3.8%
1 XLT	3404	2.8%
1 XL	3306	2.8%
6 XL	3148	2.6%
5 XLT	1753	1.5%
XLT	1068	0.9%
7 XL	648	0.5%
6 XLT	533	0.4%
8 XL	319	0.3%
7 XLT	250	0.2%
LT	157	0.1%
XL	130	0.1%
13-16	99	0.1%
Other values (35)	716	0.6%

@manelbutterfly

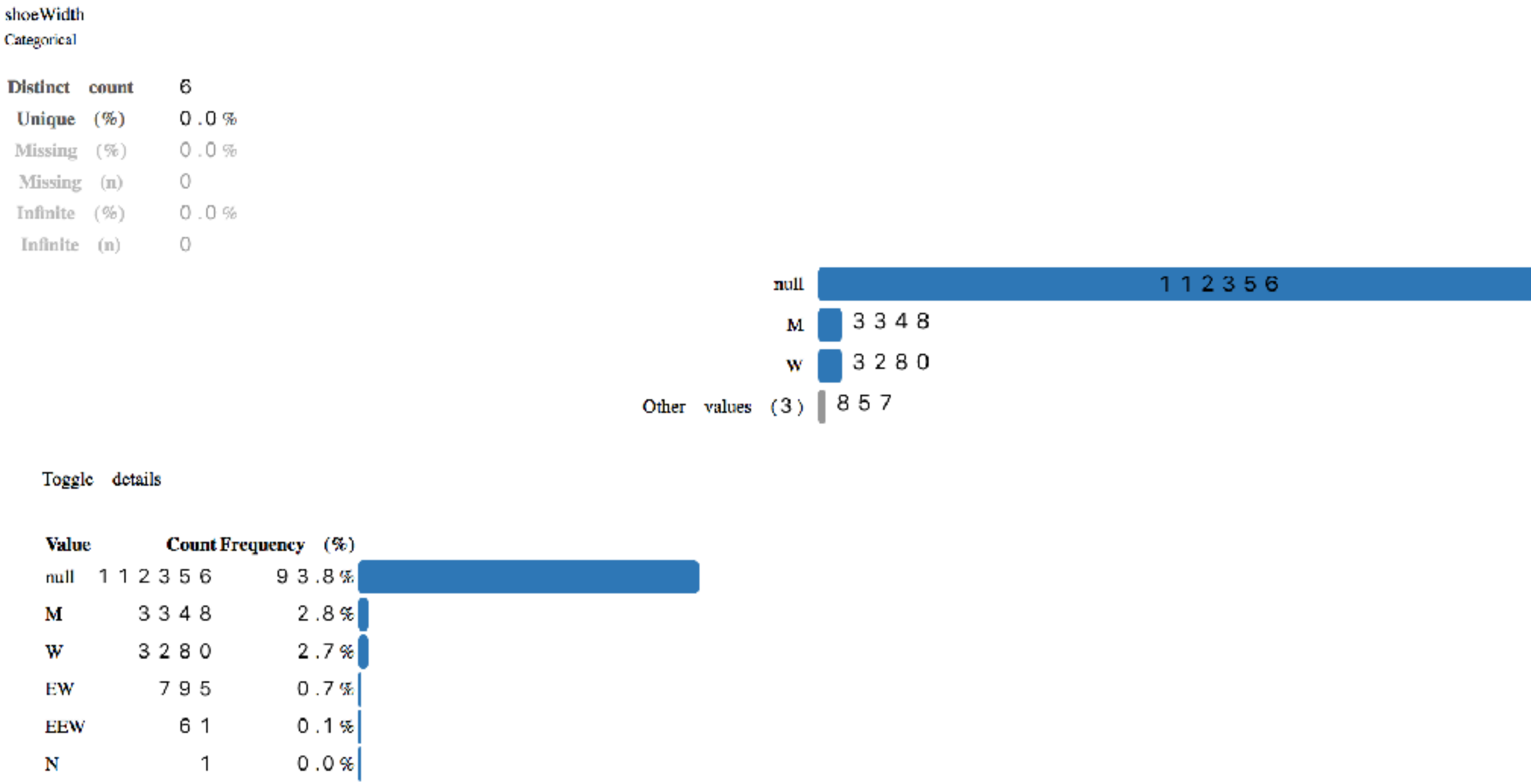


shortLength
Categorical

Distinct count	7
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Toggle details

Value	Count	Frequency (%)
null	117755	98.3%
Reg	1142	1.0%
Long	401	0.3%
LONG	253	0.2%
BIG	133	0.1%
TALL	107	0.1%
SHORT	50	0.0%



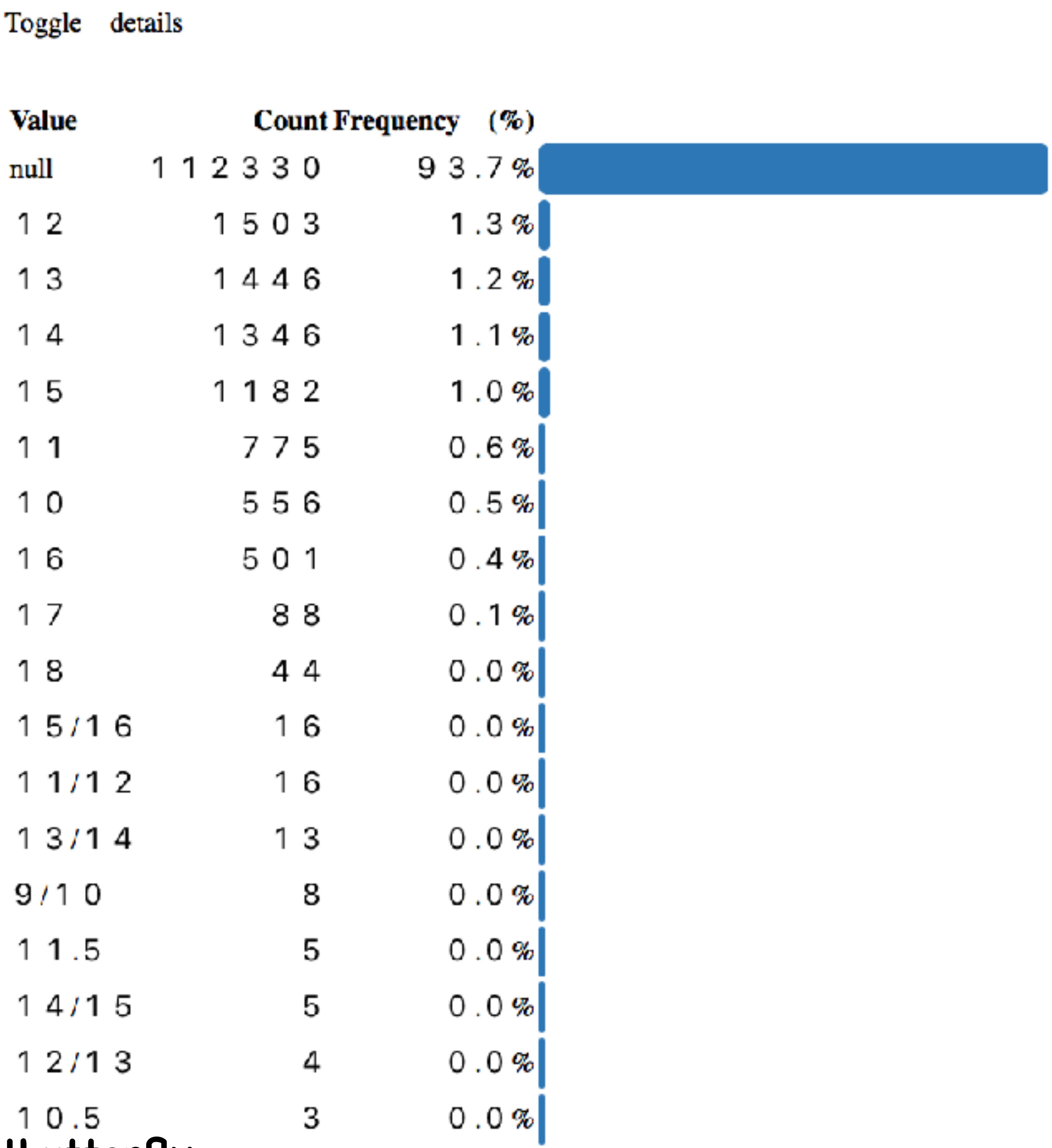
shoeWidth

Categorical

Distinct count	6
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0

Toggle details

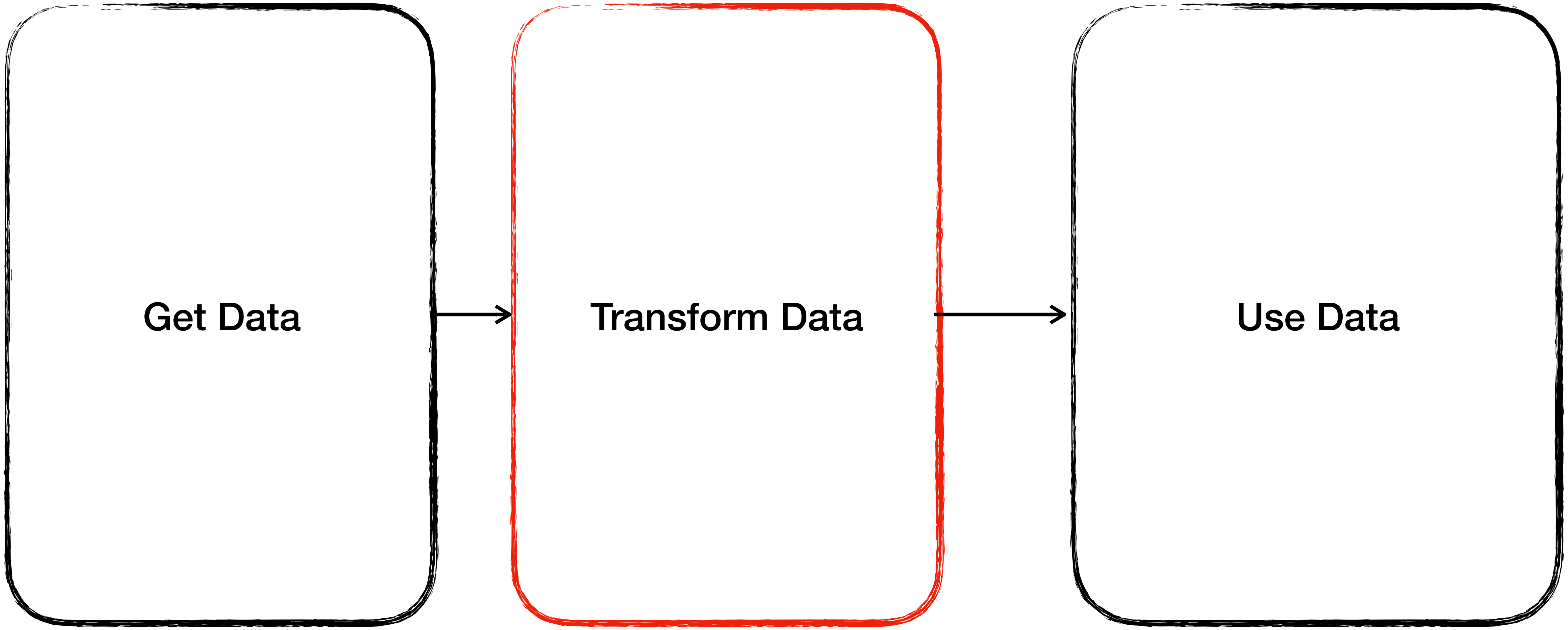
Value	Count	Frequency (%)
null	112356	93.8%
M	3348	2.8%
W	3280	2.7%
EW	795	0.7%
EEW	61	0.1%
N	1	0.0%



Toggle details

Value	Count	Frequency (%)
null	112330	93.7%
12	1503	1.3%
13	1446	1.2%
14	1346	1.1%
15	1182	1.0%
11	775	0.6%
10	556	0.5%
16	501	0.4%
17	88	0.1%
18	44	0.0%
15/16	16	0.0%
11/12	16	0.0%
13/14	13	0.0%
9/10	8	0.0%
11.5	5	0.0%
14/15	5	0.0%
12/13	4	0.0%
10.5	3	0.0%

What We Do?



Get Data

Transform Data

Use Data

@manelbutterfly

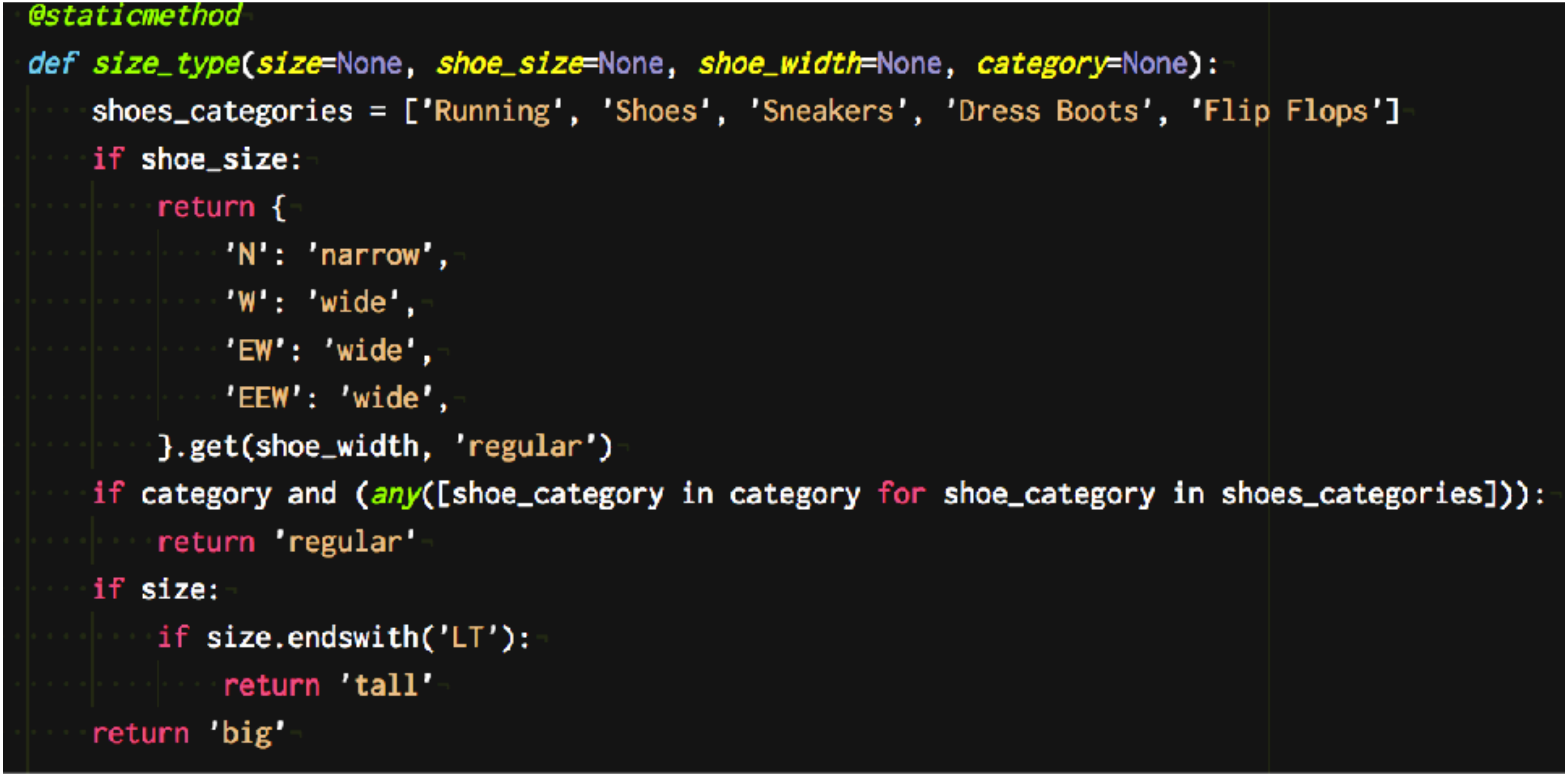
Testing Rules ٩✌°◡°✌۶

Write Ingestion

Write UDF functions

Test them

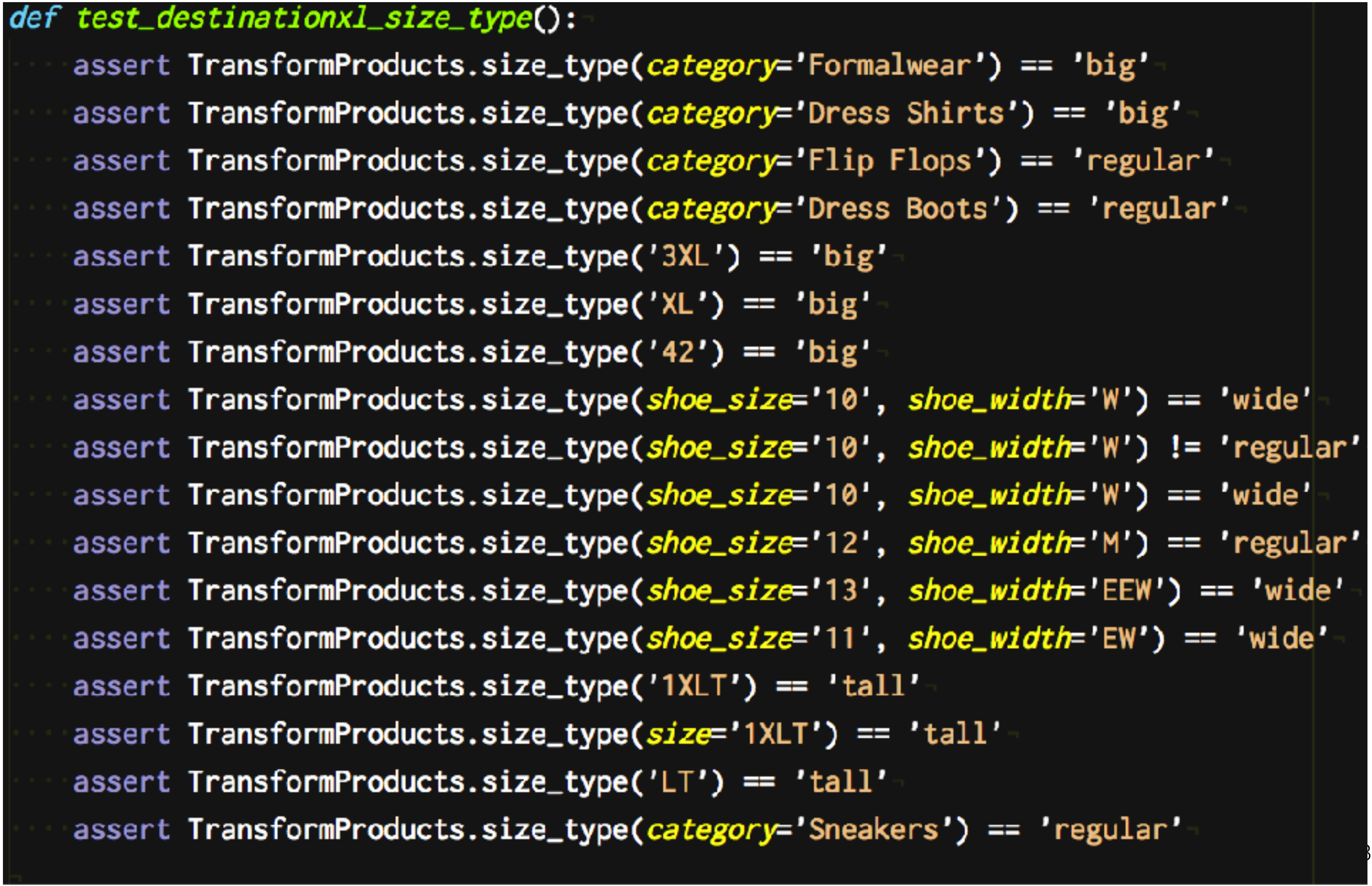
Testing Rules ୧✌°◡°✌୨



```python
@staticmethod
def size_type(size=None, shoe_size=None, shoe_width=None, category=None):
    shoes_categories = ['Running', 'Shoes', 'Sneakers', 'Dress Boots', 'Flip Flops']
    if shoe_size:
        return {
            'N': 'narrow',
            'W': 'wide',
            'EW': 'wide',
            'EEW': 'wide',
        }.get(shoe_width, 'regular')
    if category and (any([shoe_category in category for shoe_category in shoes_categories])):
        return 'regular'
    if size:
        if size.endswith('LT'):
            return 'tall'
    return 'big'
```

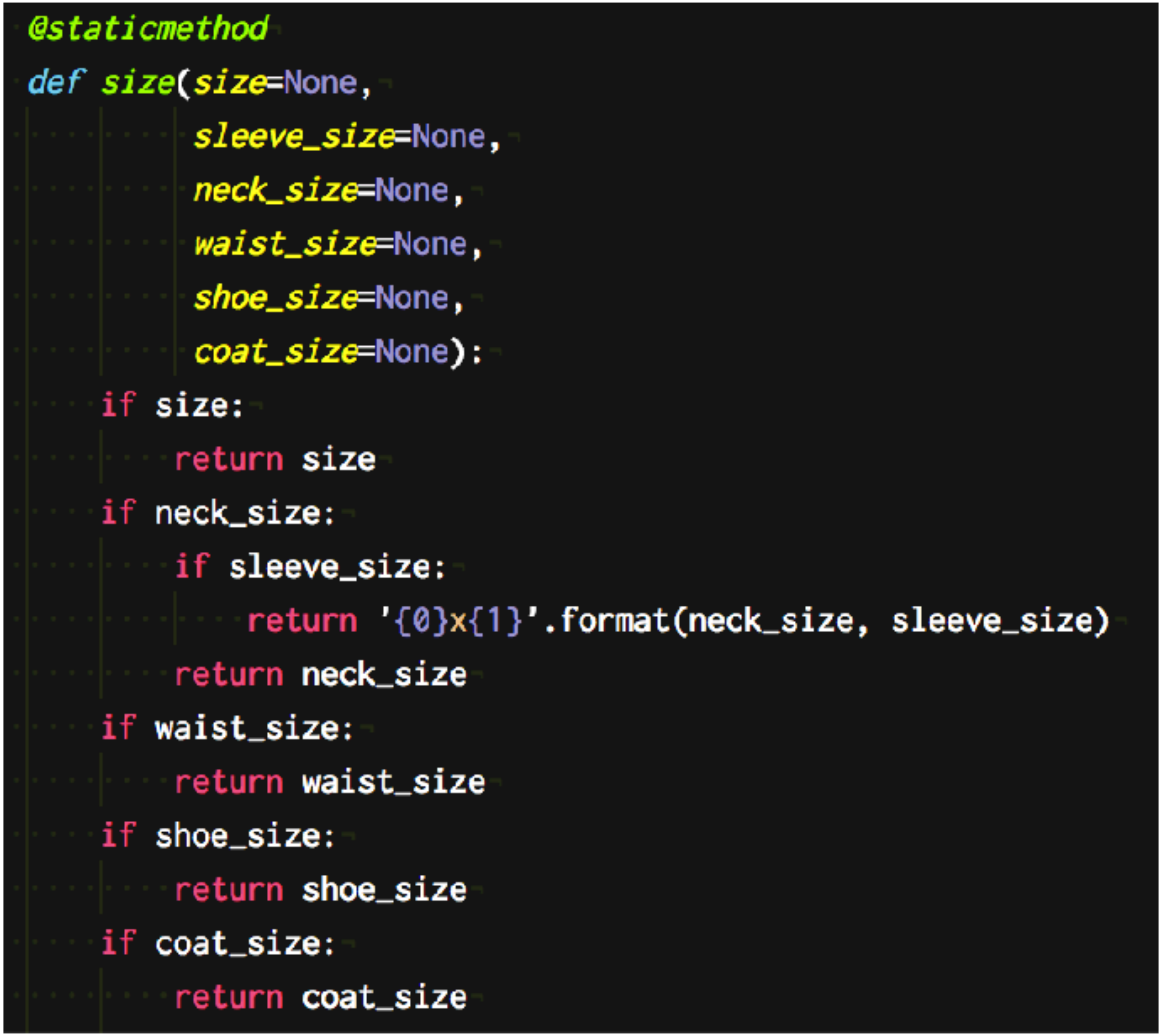
@manelbutterfly

Testing Rules ୧✌°‿°✌୨



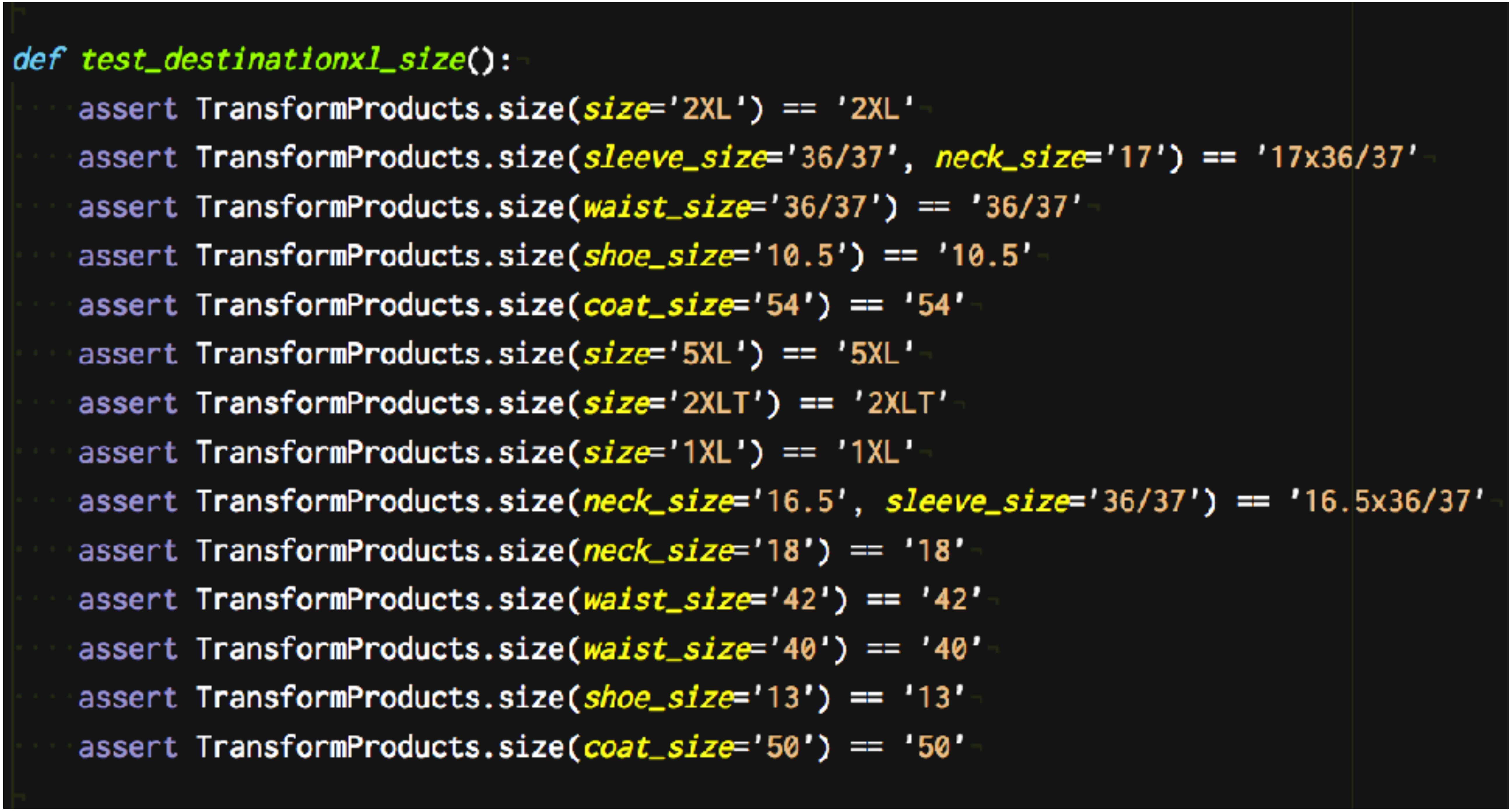
```
def test_destinationxl_size_type():
    assert TransformProducts.size_type(category='Formalwear') == 'big'
    assert TransformProducts.size_type(category='Dress Shirts') == 'big'
    assert TransformProducts.size_type(category='Flip Flops') == 'regular'
    assert TransformProducts.size_type(category='Dress Boots') == 'regular'
    assert TransformProducts.size_type('3XL') == 'big'
    assert TransformProducts.size_type('XL') == 'big'
    assert TransformProducts.size_type('42') == 'big'
    assert TransformProducts.size_type(shoe_size='10', shoe_width='W') == 'wide'
    assert TransformProducts.size_type(shoe_size='10', shoe_width='W') != 'regular'
    assert TransformProducts.size_type(shoe_size='10', shoe_width='W') == 'wide'
    assert TransformProducts.size_type(shoe_size='12', shoe_width='M') == 'regular'
    assert TransformProducts.size_type(shoe_size='13', shoe_width='EEW') == 'wide'
    assert TransformProducts.size_type(shoe_size='11', shoe_width='EW') == 'wide'
    assert TransformProducts.size_type('1XLT') == 'tall'
    assert TransformProducts.size_type(size='1XLT') == 'tall'
    assert TransformProducts.size_type('LT') == 'tall'
    assert TransformProducts.size_type(category='Sneakers') == 'regular'
```

Testing Rules ୧✌°◡°✌୨



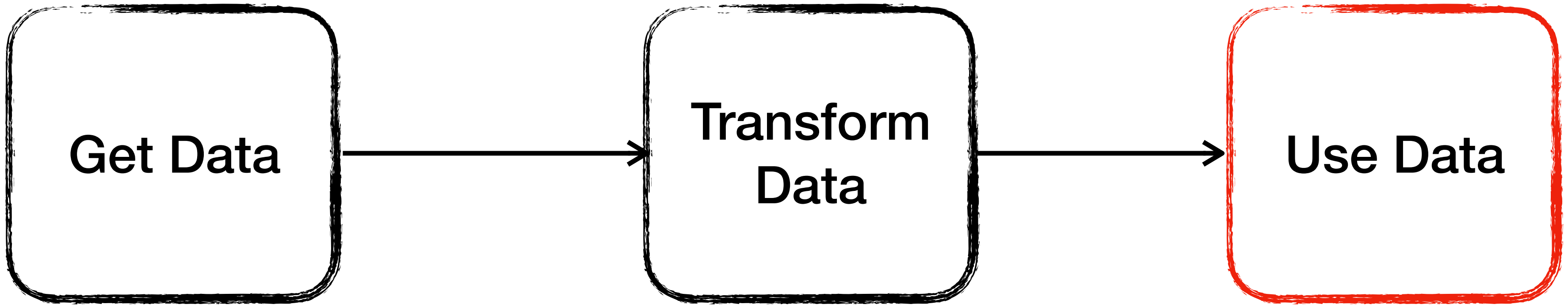
```python
@staticmethod
def size(size=None,
         sleeve_size=None,
         neck_size=None,
         waist_size=None,
         shoe_size=None,
         coat_size=None):
    if size:
        return size
    if neck_size:
        if sleeve_size:
            return '{0}x{1}'.format(neck_size, sleeve_size)
        return neck_size
    if waist_size:
        return waist_size
    if shoe_size:
        return shoe_size
    if coat_size:
        return coat_size
```

Testing Rules ٩✌°◡°✌۶



```python
def test_destinationxl_size():
    assert TransformProducts.size(size='2XL') == '2XL'
    assert TransformProducts.size(sleeve_size='36/37', neck_size='17') == '17x36/37'
    assert TransformProducts.size(waist_size='36/37') == '36/37'
    assert TransformProducts.size(shoe_size='10.5') == '10.5'
    assert TransformProducts.size(coat_size='54') == '54'
    assert TransformProducts.size(size='5XL') == '5XL'
    assert TransformProducts.size(size='2XLT') == '2XLT'
    assert TransformProducts.size(size='1XL') == '1XL'
    assert TransformProducts.size(neck_size='16.5', sleeve_size='36/37') == '16.5x36/37'
    assert TransformProducts.size(neck_size='18') == '18'
    assert TransformProducts.size(waist_size='42') == '42'
    assert TransformProducts.size(waist_size='40') == '40'
    assert TransformProducts.size(shoe_size='13') == '13'
    assert TransformProducts.size(coat_size='50') == '50'
```

What We Do?


Get Data

Transform Data

Use Data

FIT PREDICTOR

@manelbutterfly

“Maybe stories are just data with a soul.”

-Brené Brown