Overview

Dataset statistics

Number of variables	8
Number of observations	1167
Missing cells	1651
Missing cells (%)	17.7%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	73.1 KiB
Average record size in memory	64.1 B

Variable types

Unsupported	2
Numeric	1
Categorical	2
Text	2
Boolean	1

Alerts

barcode is highly overall correlated with categoryCode and 1 other fields (categoryCode, topBrand)	High correlation
category is highly overall correlated with categoryCode	High correlation
categoryCode is highly overall correlated with barcode and 2 other fields (barcode, category, topBrand)	High correlation
topBrand is highly overall correlated with barcode and 1 other fields (barcode, categoryCode)	High correlation
categoryCode is highly imbalanced (56.0%)	Imbalance
topBrand is highly imbalanced (68.9%)	Imbalance
category has 155 (13.3%) missing values	Missing
categoryCode has 650 (55.7%) missing values	Missing
topBrand has 612 (52.4%) missing values	Missing
brandCode has 234 (20.1%) missing values	Missing
_id is an unsupported type, check if it needs cleaning or further analysis	Unsupported
cpg is an unsupported type, check if it needs cleaning or further analysis	Unsupported

Reproduction

Analysis started	2024-10-09 04:07:50.233114
Analysis finished	2024-10-09 04:07:57.471754
Duration	7.24 seconds
Software version	ydata-profiling vv4.10.0 (https://github.com/ydataai/ydata-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22Brands%20Profile%20Report%22%2C%20%22dataset%22%3A%20%7B%22description%22%3A%20%22%22%2C%20%22creator%22%3A%20%22%22%2C%20%22author%22%3A%20%22%2

Variables

Select Columns

~~_id~~
~~Unsupported~~

REJECTED UNSUPPORTED

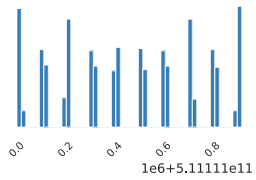
Missing	0
Missing (%)	0.0%
Memory size	9.2 KiB

barcode

Real number ()

HIGH CORRELATION (This variable has a high overall correlation with 2 fields: categoryCode, topBrand)

Distinct	1160
Distinct (%)	99.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5.1111146×10^{11}
Minimum	5.11111×10^{11}
Maximum	5.1111192×10^{11}
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	9.2 KiB

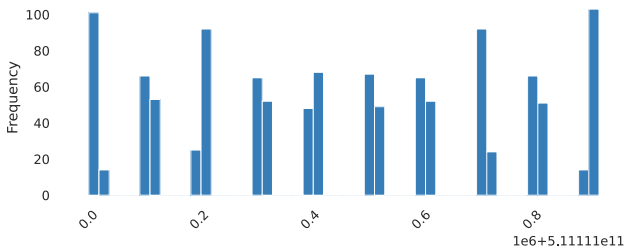


Quantile statistics

Minimum	5.11111×10^{11}
5-th percentile	5.1111101×10^{11}
Q1	5.1111121×10^{11}
median	5.1111142×10^{11}
Q3	5.1111171×10^{11}
95-th percentile	5.1111191×10^{11}
Maximum	5.1111192×10^{11}
Range	919636
Interquartile range (IQR)	500000

Descriptive statistics

Standard deviation	287449.75
Coefficient of variation (CV)	5.624013×10^{-7}
Kurtosis	-1.2258366
Mean	5.1111146×10^{11}
Median Absolute Deviation (MAD)	280214
Skewness	0.0036519198
Sum	5.9646707×10^{14}
Variance	8.2627358×10^{10}
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
$5.111113051 \times 10^{11}$	2	0.2%
$5.111115048 \times 10^{11}$	2	0.2%
$5.111117041 \times 10^{11}$	2	0.2%
$5.111110048 \times 10^{11}$	2	0.2%
$5.111115041 \times 10^{11}$	2	0.2%
$5.111116051 \times 10^{11}$	2	0.2%
$5.111112049 \times 10^{11}$	2	0.2%
$5.111119033 \times 10^{11}$	1	0.1%
$5.111112018 \times 10^{11}$	1	0.1%
$5.111118165 \times 10^{11}$	1	0.1%
Other values (1150)	1150	98.5%

Value	Count	Frequency (%)
$5.111110002 \times 10^{11}$	1	0.1%
$5.111110002 \times 10^{11}$	1	0.1%
$5.111110004 \times 10^{11}$	1	0.1%
$5.111110004 \times 10^{11}$	1	0.1%
$5.111110005 \times 10^{11}$	1	0.1%
$5.111110007 \times 10^{11}$	1	0.1%
$5.111110007 \times 10^{11}$	1	0.1%
$5.111110009 \times 10^{11}$	1	0.1%
$5.11111001 \times 10^{11}$	1	0.1%
$5.11111001 \times 10^{11}$	1	0.1%

Value	Count	Frequency (%)
$5.111119198 \times 10^{11}$	1	0.1%
$5.111119198 \times 10^{11}$	1	0.1%
$5.111119197 \times 10^{11}$	1	0.1%
$5.111119195 \times 10^{11}$	1	0.1%
$5.111119195 \times 10^{11}$	1	0.1%
$5.111119193 \times 10^{11}$	1	0.1%
$5.111119192 \times 10^{11}$	1	0.1%
$5.111119192 \times 10^{11}$	1	0.1%
$5.11111919 \times 10^{11}$	1	0.1%
$5.111119189 \times 10^{11}$	1	0.1%

category

Categorical

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: categoryCode) MISSING

Distinct	23
Distinct (%)	2.3%
Missing	155
Missing (%)	13.3%
Memory size	9.2 KiB

Length

Max length	27
Median length	22
Mean length	9.7628458
Min length	4

Characters and Unicode

Total characters	9880	
Distinct characters	38	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1	?
Unique (%)	0.1%	

Sample

1st row	Baking
2nd row	Beverages
3rd row	Baking
4th row	Baking
5th row	Candy & Sweets

Common Values

Value	Count	Frequency (%)
Baking	369	31.6%
Beer Wine Spirits	90	7.7%
Snacks	75	6.4%
Candy & Sweets	71	6.1%
Beverages	63	5.4%
Magazines	44	3.8%
Health & Wellness	44	3.8%
Breakfast & Cereal	40	3.4%
Grocery	39	3.3%
Dairy	33	2.8%
Other values (13)	144	12.3%
(Missing)	155	13.3%

Length

Value	Count	Frequency (%)
baking	369	22.1%
	216	12.9%
beer	90	5.4%
wine	90	5.4%
spirits	90	5.4%
snacks	75	4.5%
candy	71	4.3%
sweets	71	4.3%
beverages	63	3.8%
magazines	44	2.6%
Other values (25)	489	29.3%

Most occurring characters

Value	Count	Frequency (%)
e	1153	11.7%
a	1013	10.3%
n	839	8.5%
i	765	7.7%
	656	6.6%
B	605	6.1%
s	580	5.9%
r	542	5.5%
k	489	4.9%
g	487	4.9%
Other values (28)	2751	27.8%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	9880	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
e	1153	11.7%
a	1013	10.3%
n	839	8.5%
i	765	7.7%
	656	6.6%
B	605	6.1%
s	580	5.9%
r	542	5.5%
k	489	4.9%
g	487	4.9%
Other values (28)	2751	27.8%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	9880	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
e	1153	11.7%
a	1013	10.3%
n	839	8.5%
i	765	7.7%
	656	6.6%
B	605	6.1%
s	580	5.9%
r	542	5.5%
k	489	4.9%
g	487	4.9%
Other values (28)	2751	27.8%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	9880	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
e	1153	11.7%
a	1013	10.3%
n	839	8.5%
i	765	7.7%
	656	6.6%
B	605	6.1%
s	580	5.9%
r	542	5.5%
k	489	4.9%
g	487	4.9%
Other values (28)	2751	27.8%

categoryCode

Categorical

HIGH CORRELATION (This variable has a high overall correlation with 3 fields: barcode, category, topBrand)	IMBALANCE	MISSING

Distinct	14
Distinct (%)	2.7%
Missing	650
Missing (%)	55.7%
Memory size	9.2 KiB

Length

Max length	29
Median length	6
Mean length	8.9922631
Min length	4

Characters and Unicode

Total characters	4649	
Distinct characters	24	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	**?**
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	**?**
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	**?**

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	4	**?**
Unique (%)	0.8%	

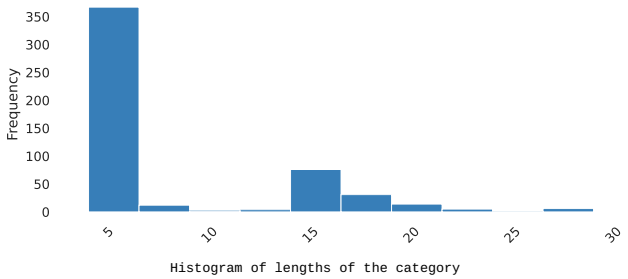
Sample

1st row	BAKING
2nd row	BEVERAGES
3rd row	BAKING
4th row	BAKING
5th row	CANDY_AND_SWEETS

Common Values

Value	Count	Frequency (%)
BAKING	359	30.8%
CANDY_AND_SWEETS	71	6.1%
BEER_WINE_SPIRITS	31	2.7%
HEALTHY_AND_WELLNESS	14	1.2%
GROCERY	11	0.9%
BABY	7	0.6%
CLEANING_AND_HOME_IMPROVEMENT	6	0.5%
BREAD_AND_BAKERY	5	0.4%
DAIRY_AND_REFRIGERATED	5	0.4%
PERSONAL_CARE	4	0.3%
Other values (4)	4	0.3%
(Missing)	650	55.7%

Length



Value	Count	Frequency (%)
baking	359	69.4%
candy_and_sweets	71	13.7%
beer_wine_spirits	31	6.0%
healthy_and_wellness	14	2.7%
grocery	11	2.1%
baby	7	1.4%
cleaning_and_home_improvement	6	1.2%
bread_and_bakery	5	1.0%
dairy_and_refrigerated	5	1.0%
personal_care	4	0.8%
Other values (4)	4	0.8%

Most occurring characters

Value	Count	Frequency (%)
N	600	12.9%
A	589	12.7%
I	475	10.2%
B	415	8.9%
G	383	8.2%
K	364	7.8%
E	350	7.5%
_	274	5.9%
S	238	5.1%
D	188	4.0%
Other values (14)	773	16.6%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	4649	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
N	600	12.9%
A	589	12.7%
I	475	10.2%
B	415	8.9%
G	383	8.2%
K	364	7.8%
E	350	7.5%
_	274	5.9%
S	238	5.1%
D	188	4.0%
Other values (14)	773	16.6%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	4649	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
N	600	12.9%
A	589	12.7%
I	475	10.2%
B	415	8.9%
G	383	8.2%
K	364	7.8%
E	350	7.5%
_	274	5.9%
S	238	5.1%
D	188	4.0%
Other values (14)	773	16.6%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	4649	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
N	600	12.9%
A	589	12.7%
I	475	10.2%
B	415	8.9%
G	383	8.2%
K	364	7.8%
E	350	7.5%
_	274	5.9%
S	238	5.1%
D	188	4.0%
Other values (14)	773	16.6%

cpg

Unsupported

REJECTED UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	9.2 KiB

name

Text

Distinct	1156
Distinct (%)	99.1%
Missing	0
Missing (%)	0.0%
Memory size	9.2 KiB

Length

Max length	57
Median length	46
Mean length	17.658098
Min length	1

Characters and Unicode

Total characters	20607	
Distinct characters	80	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1145	?
Unique (%)	98.1%	

Sample

1st row	test brand @1612366101024
2nd row	Starbucks
3rd row	test brand @1612366146176
4th row	test brand @1612366146051
5th row	test brand @1612366146827

Value	Count	Frequency (%)
brand	433	15.2%
test	432	15.2%
magazine	34	1.2%
	22	0.8%
a	15	0.5%
one	13	0.5%
special	12	0.4%
day®	12	0.4%
edition	11	0.4%
and	10	0.4%
Other values (1395)	1848	65.0%

Most occurring characters

Value	Count	Frequency (%)
	1692	8.2%
e	1234	6.0%
t	1216	5.9%
a	1080	5.2%
1	966	4.7%
r	908	4.4%
s	817	4.0%
n	801	3.9%
6	746	3.6%
0	664	3.2%
Other values (70)	10483	50.9%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	20607	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
	1692	8.2%
e	1234	6.0%
t	1216	5.9%
a	1080	5.2%
1	966	4.7%
r	908	4.4%
s	817	4.0%
n	801	3.9%
6	746	3.6%
0	664	3.2%
Other values (70)	10483	50.9%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	20607	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
	1692	8.2%
e	1234	6.0%
t	1216	5.9%
a	1080	5.2%
1	966	4.7%
r	908	4.4%
s	817	4.0%
n	801	3.9%
6	746	3.6%
0	664	3.2%
Other values (70)	10483	50.9%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	20607	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
	1692	8.2%
e	1234	6.0%
t	1216	5.9%
a	1080	5.2%
1	966	4.7%
r	908	4.4%
s	817	4.0%
n	801	3.9%
6	746	3.6%
0	664	3.2%
Other values (70)	10483	50.9%

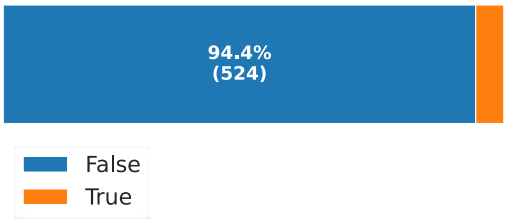
topBrand

Boolean

HIGH CORRELATION (This variable has a high overall correlation with 2 fields: barcode, categoryCode) IMBALANCE MISSING

Distinct	2
Distinct (%)	0.4%
Missing	612
Missing (%)	52.4%
Memory size	9.2 KiB

Value	Count	Frequency (%)
False	524	44.9%
True	31	2.7%
(Missing)	612	52.4%

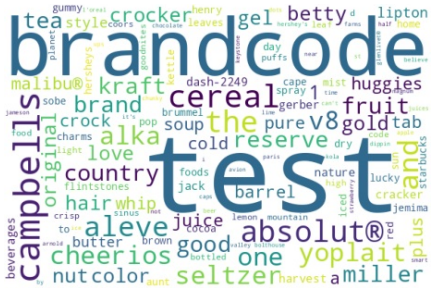

94.4%
(524)
False
True

brandCode

Text

MISSING

Distinct	897
Distinct (%)	96.1%
Missing	234
Missing (%)	20.1%
Memory size	9.2 KiB



Length

Max length	46
Median length	40
Mean length	17.906752
Min length	0

Characters and Unicode

Total characters	16707	
Distinct characters	65	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

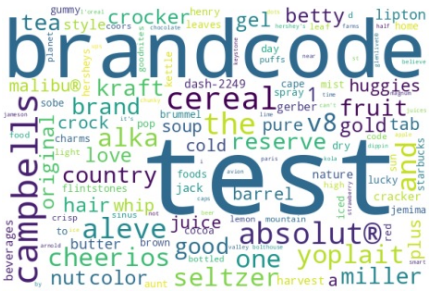
Unique

Unique	894	?
Unique (%)	95.8%	

Sample

1st row	STARBUCKS
2nd row	TEST BRANDCODE @1612366146176
3rd row	TEST BRANDCODE @1612366146051
4th row	TEST BRANDCODE @1612366146827
5th row	TEST BRANDCODE @1612366146091

Value	Count	Frequency (%)
test	359	17.5%
brandcode	357	17.4%
	14	0.7%
campbells	10	0.5%
cereal	8	0.4%
the	7	0.3%
absolut®	7	0.3%
and	6	0.3%
aleve	6	0.3%
v8	5	0.2%
Other values (1094)	1273	62.0%



Most occurring characters

Value	Count	Frequency (%)
E	1285	7.7%
	1172	7.0%
1	1089	6.5%
T	1025	6.1%
D	862	5.2%
A	788	4.7%
S	751	4.5%
R	714	4.3%
O	691	4.1%
6	636	3.8%
Other values (55)	7694	46.1%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	16707	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
E	1285	7.7%
	1172	7.0%
1	1089	6.5%
T	1025	6.1%
D	862	5.2%
A	788	4.7%
S	751	4.5%
R	714	4.3%
O	691	4.1%
6	636	3.8%
Other values (55)	7694	46.1%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	16707	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
E	1285	7.7%
	1172	7.0%
1	1089	6.5%
T	1025	6.1%
D	862	5.2%
A	788	4.7%
S	751	4.5%
R	714	4.3%
O	691	4.1%
6	636	3.8%
Other values (55)	7694	46.1%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	16707	100.0%

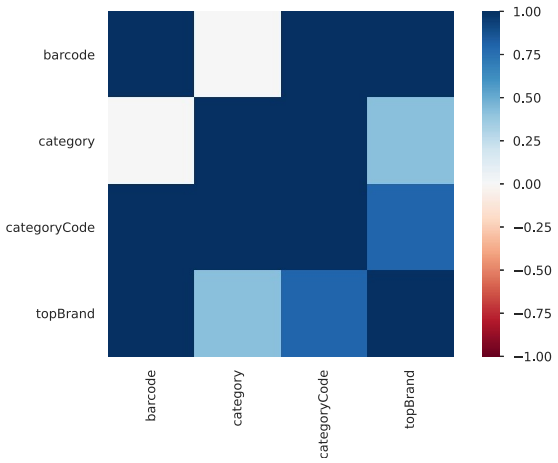
Most frequent character per block

(unknown)

Value	Count	Frequency (%)
E	1285	7.7%
	1172	7.0%
1	1089	6.5%
T	1025	6.1%
D	862	5.2%
A	788	4.7%
S	751	4.5%
R	714	4.3%
O	691	4.1%
6	636	3.8%
Other values (55)	7694	46.1%

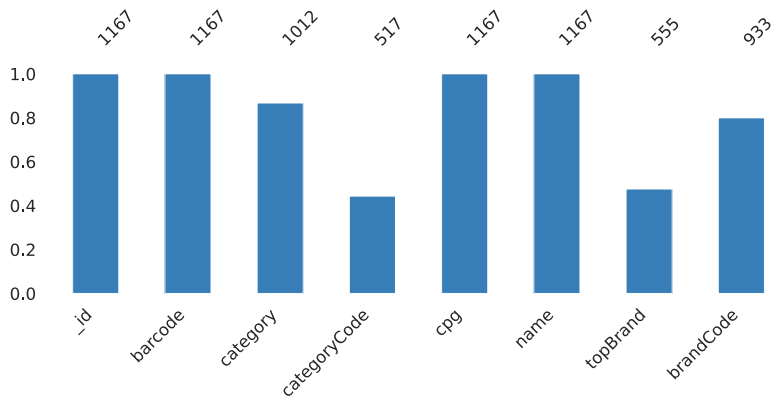
Interactions

Correlations

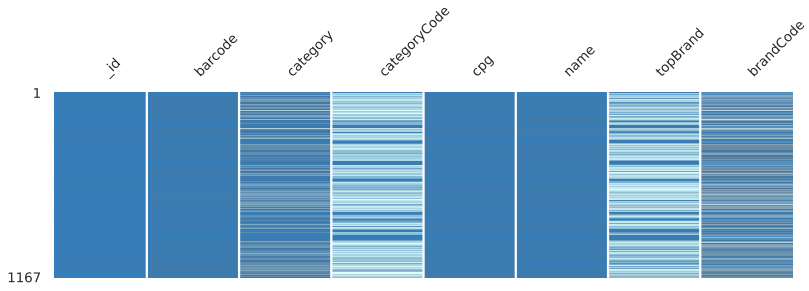


	barcode	category	categoryCode	topBrand
barcode	1.000	0.000	1.000	1.000
category	0.000	1.000	1.000	0.414
categoryCode	1.000	1.000	1.000	0.792
topBrand	1.000	0.414	0.792	1.000

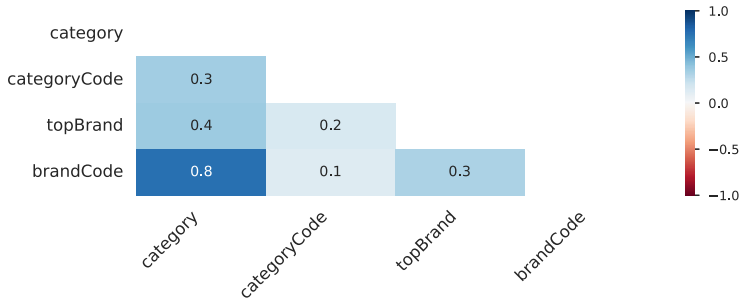
Missing values



A simple visualization of nullity by column.



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.



The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.

Sample

	_id	barcode	category	categoryCode	cpg	name	topBrand	brandCo
0	601ac115be37ce2ead437551	511111019862	Baking	BAKING	DBRef('Cogs', ObjectId('601ac114be37ce2ead437550'))	test brand @1612366101024	False	NaN
1	601c5460be37ce2ead43755f	511111519928	Beverages	BEVERAGES	DBRef('Cogs', ObjectId('5332f5fbe4b03c9a25efd0ba'))	Starbucks	False	STARBUC
2	601ac142be37ce2ead43755d	511111819905	Baking	BAKING	DBRef('Cogs', ObjectId('601ac142be37ce2ead437559'))	test brand @1612366146176	False	TEST BR
3	601ac142be37ce2ead43755a	511111519874	Baking	BAKING	DBRef('Cogs', ObjectId('601ac142be37ce2ead437559'))	test brand @1612366146051	False	TEST BR
4	601ac142be37ce2ead43755e	511111319917	Candy & Sweets	CANDY_AND_SWEETS	DBRef('Cogs', ObjectId('5332fa12e4b03c9a25efd1e7'))	test brand @1612366146827	False	TEST BR
5	601ac142be37ce2ead43755b	511111719885	Baking	BAKING	DBRef('Cogs', ObjectId('601ac142be37ce2ead437559'))	test brand @1612366146091	False	TEST BR
6	601ac142be37ce2ead43755c	511111219897	Baking	BAKING	DBRef('Cogs', ObjectId('601ac142be37ce2ead437559'))	test brand @1612366146133	False	TEST BR
7	5cdad0f5166eb33eb7ce0faa	511111104810	Condiments & Sauces	NaN	DBRef('Cogs', ObjectId('559c2234e4b06aca36af13c6'))	J.L. Kraft	NaN	J.L. KR
8	5ab15636e4b0be0a89bb0b07	511111504412	Canned Goods & Soups	NaN	DBRef('Cogs', ObjectId('5a734034e4b0d58f376be874'))	Campbell's Home Style	False	CAMPBEL
9	5c408e8bcd244a1fdb47aee7	511111504788	Baking	NaN	DBRef('Cogs', ObjectId('59ba6f1ce4b092b29c167346'))	test	NaN	TEST

	_id	barcode	category	categoryCode	cpg	name	topBrand	brandC
1157	5332fa75e4b03c9a25efd221	511111303015	NaN	NaN	DBRef('Cpgs', ObjectId('5332f5ebe4b03c9a25efd0a8'))	DASANI	NaN	NaN
1158	5f628215be37ce78e6e2efab	511111716648	Baking	BAKING	DBRef('Cogs', ObjectId('5f628214be37ce78e6e2efaa'))	test brand @1600291349042	NaN	TEST E
1159	585a96cbe4b03e62d1ce0e88	511111501619	Beverages	NaN	DBRef('Cogs', ObjectId('5332f5fbe4b03c9a25efd0ba'))	Pepsi Max	False	
1160	5887a216e4b02187f85cdad5	511111401155	Deli	NaN	DBRef('Cogs', ObjectId('559c2234e4b06aca36af13c6'))	Claussen	False	CLAUSS
1161	5332f709e4b03c9a25efd0f2	511111403845	Beer Wine Spirits	NaN	DBRef('Cogs', ObjectId('5332f709e4b03c9a25efd0f1'))	Blue Moon	False	BLUE M
1162	5f77274dbe37ce6b592e90c0	511111116752	Baking	BAKING	DBRef('Cogs', ObjectId('5f77274dbe37ce6b592e90bf'))	test brand @1601644365844	NaN	NaN
1163	5dc1fca91dda2c0ad7da64ae	511111706328	Breakfast & Cereal	NaN	DBRef('Cogs', ObjectId('53e10d6368abd3c7065097cc'))	Dippin Dots® Cereal	NaN	DIPPIN
1164	5f494c6e04db711dd8fe87e7	511111416173	Candy & Sweets	CANDY_AND_SWEETS	DBRef('Cogs', ObjectId('5332fa12e4b03c9a25efd1e7'))	test brand @1598639215217	NaN	TEST E
1165	5a021611e4b00efe02b02a57	511111400608	Grocery	NaN	DBRef('Cogs', ObjectId('5332f5f6e4b03c9a25efd0b4'))	LIPTON TEA Leaves	False	LIPTON
1166	6026d757be37ce6369301468	511111019930	Baking	BAKING	DBRef('Cogs', ObjectId('6026d757be37ce6369301467'))	test brand @1613158231643	False	TEST E