Overview

Dataset statistics

Number of variables	12
Number of observations	550068
Missing cells	556885
Missing cells (%)	8.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	50.4 MiB
Average record size in memory	96.0 B

Variable types

NUM	6
CAT	5
BOOL	1

Warnings

Product_ID has a high cardinality: 3631 distinct values	**High cardinality**
Product_Category_2 has 173638 (31.6%) missing values	**Missing**
Product_Category_3 has 383247 (69.7%) missing values	**Missing**
Occupation has 69638 (12.7%) zeros	**Zeros**

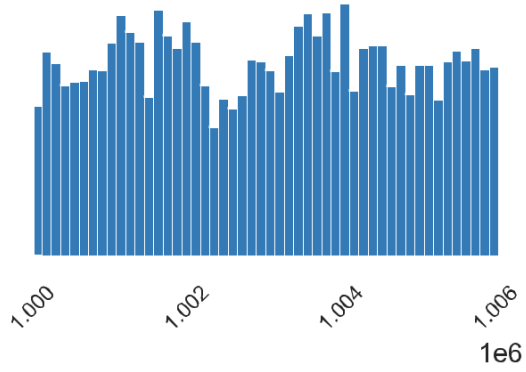
Variables

User_ID

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	5891
Distinct (%)	1.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	1003028.842
Minimum	1000001
Maximum	1006040
Zeros	0
Zeros (%)	0.0%
Memory size	4.2 MiB



Statistics

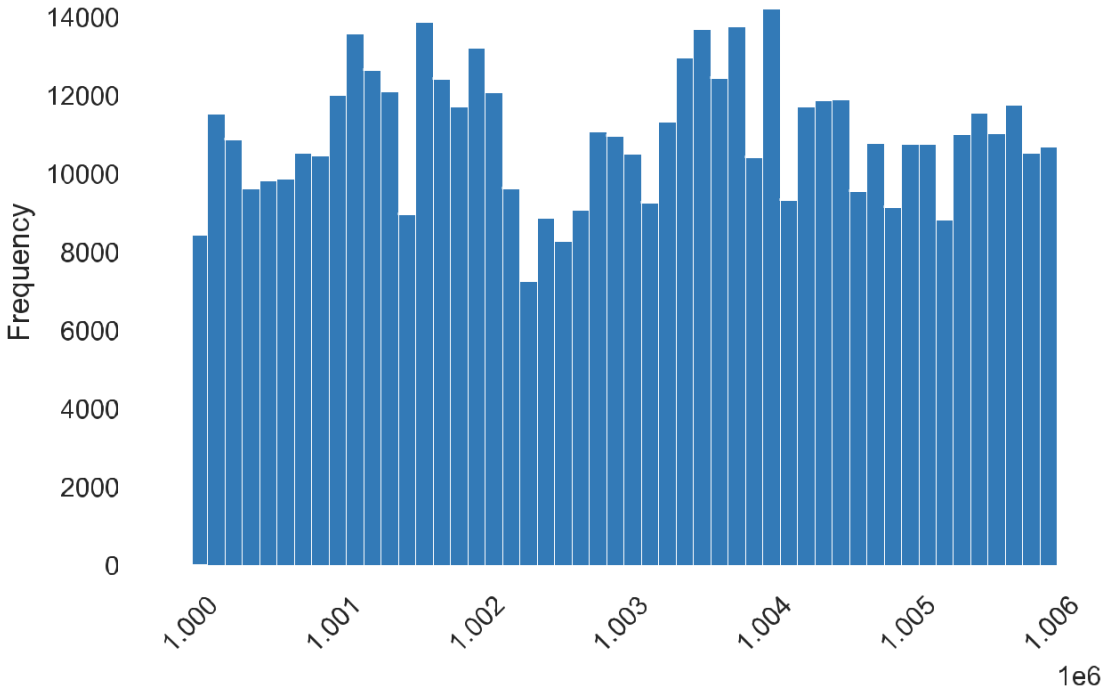
Quantile statistics

	1003077
Minimum	1000001
5-th percentile	1000329
Q1	1001516
median	
Q3	1004478
95-th percentile	1005747
Maximum	1006040
Range	6039
Interquartile range (IQR)	2962

Descriptive Statistics

Standard deviation	1727.591586
Coefficient of variation (CV)	0.001722374784
Kurtosis	-1.195500781
Mean	1003028.842
Median Absolute Deviation (MAD)	1468
Skewness	0.003065551851
Sum	5.517340693e+11
Variance	2984572.686
Monotocity	Not monotonic

Histogram



Histogram with fixed size bins (bins=50)

Common Values

Value	Count	Frequency (%)
1001680	1026	0.2%
1004277	979	0.2%
1001941	898	0.2%
1001181	862	0.2%
1000889	823	0.1%
1003618	767	0.1%
1001150	752	0.1%
1001015	740	0.1%
1005795	729	0.1%
1005831	727	0.1%
Other values (5881)	541765	98.5%

Extreme Values

Minimum 5 values

Value	Count	Frequency (%)
1000001	35	< 0.1%
1000002	77	< 0.1%
1000003	29	< 0.1%
1000004	14	< 0.1%
1000005	106	< 0.1%

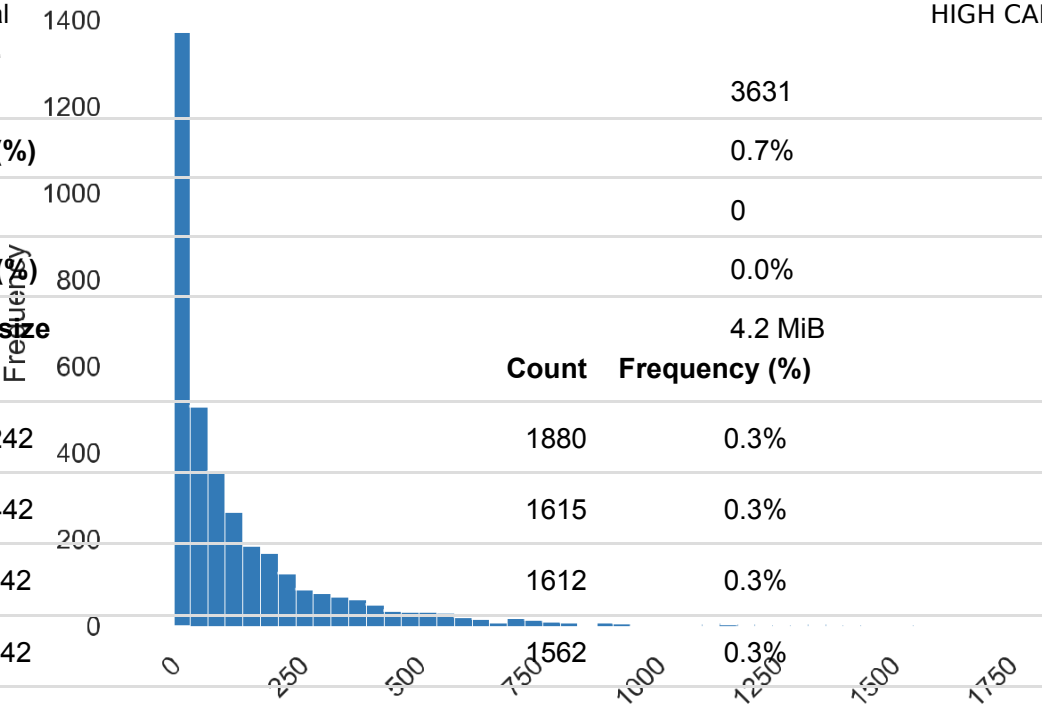
Maximum 5 values

Value	Count	Frequency (%)
1006040	180	< 0.1%
1006039	74	< 0.1%
1006038	12	< 0.1%
1006037	122	< 0.1%
1006036	514	0.1%

Product_ID

Categorical HIGH CARDINALITY

Distinct	3631
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	4.2 MiB



Value	Count	Frequency (%)
P00265242	1880	0.3%
P00025442	1615	0.3%
P00110742	1612	0.3%
P00112142	1562	0.3%
P00057642	1470	0.3%
P00184942	1440	0.3%
P00046742	1438	0.3%
P00058042	1422	0.3%
P00059442	1406	0.3%
P00145042	1406	0.3%
Other values (3621)	534817	97.2%

Frequencies

Common Values

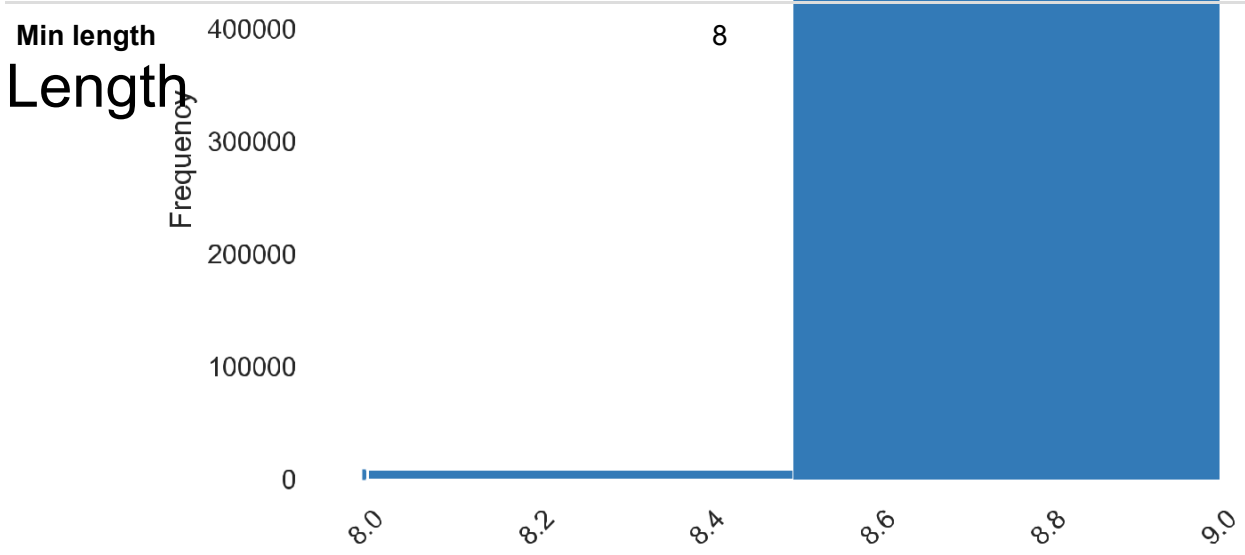
Overview

Unique	144	?
Unique (%)	< 0.1%	

Frequencies of value counts

Length

Max length	9
Median length	9
Mean length	8.982729408
Min length	8



Histogram of lengths of the category

Gender

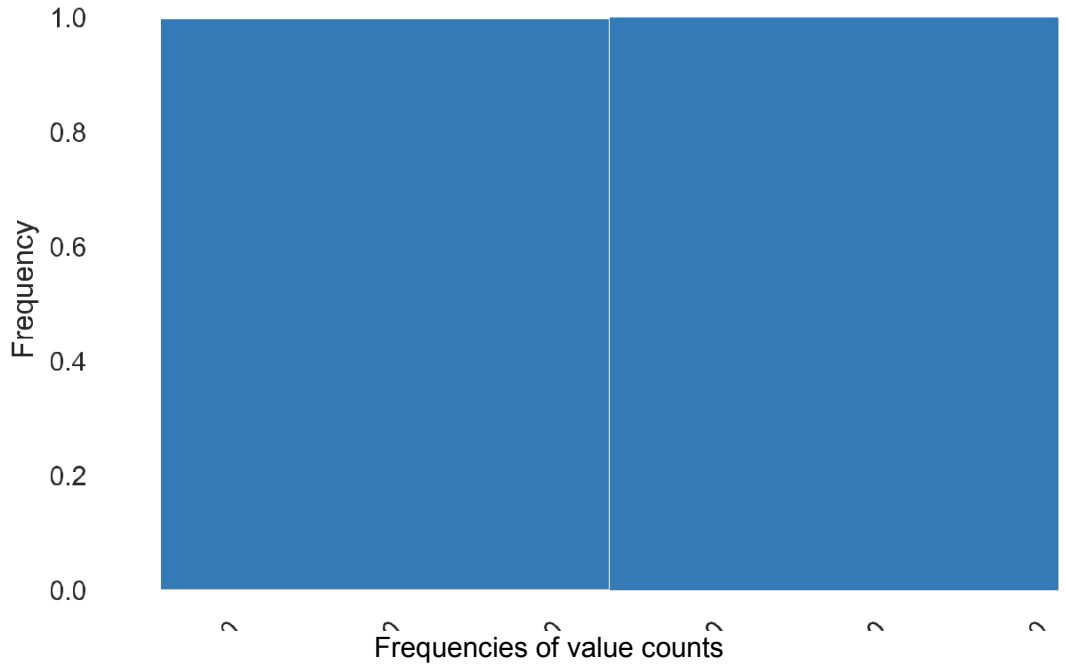
Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.2 MiB

Frequencies

Value	Count	Frequency (%)
M	414259	75.3%
F	135809	24.7%

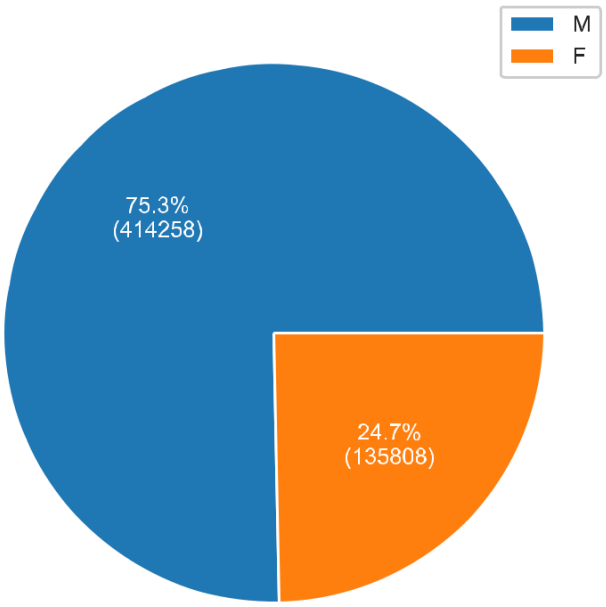
Overview
Common Values



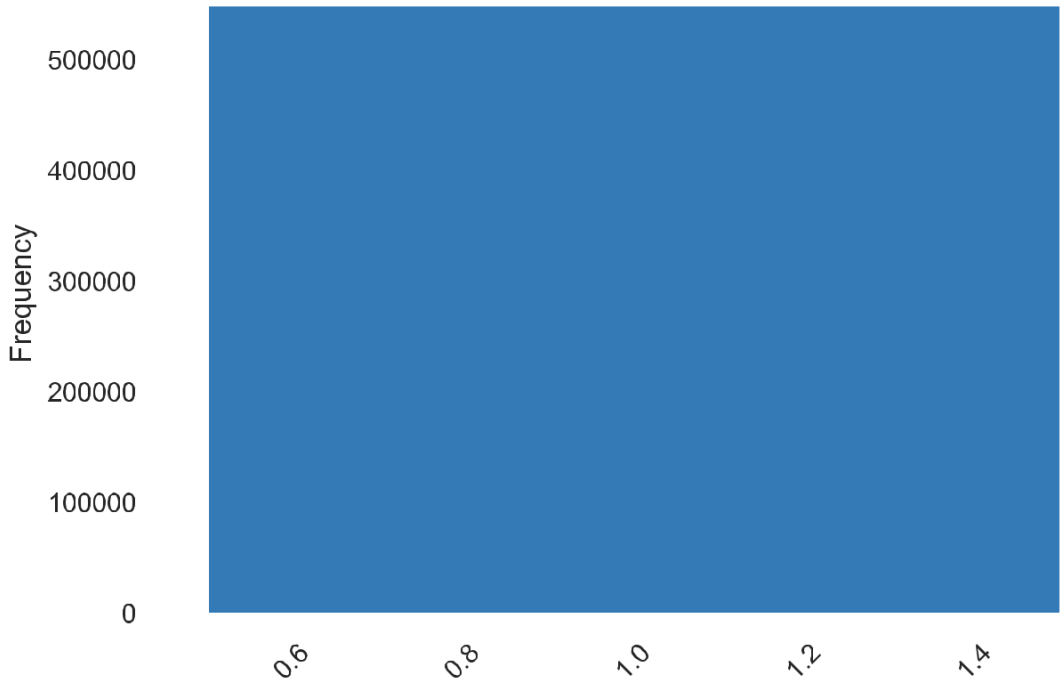
Unique

Unique	0	**?**
Unique (%)	0.0%	

Chart



Length



Max length	1
Median length	1
Mean length	1
Min length	1

Histogram of lengths of the category

Age

Categorical

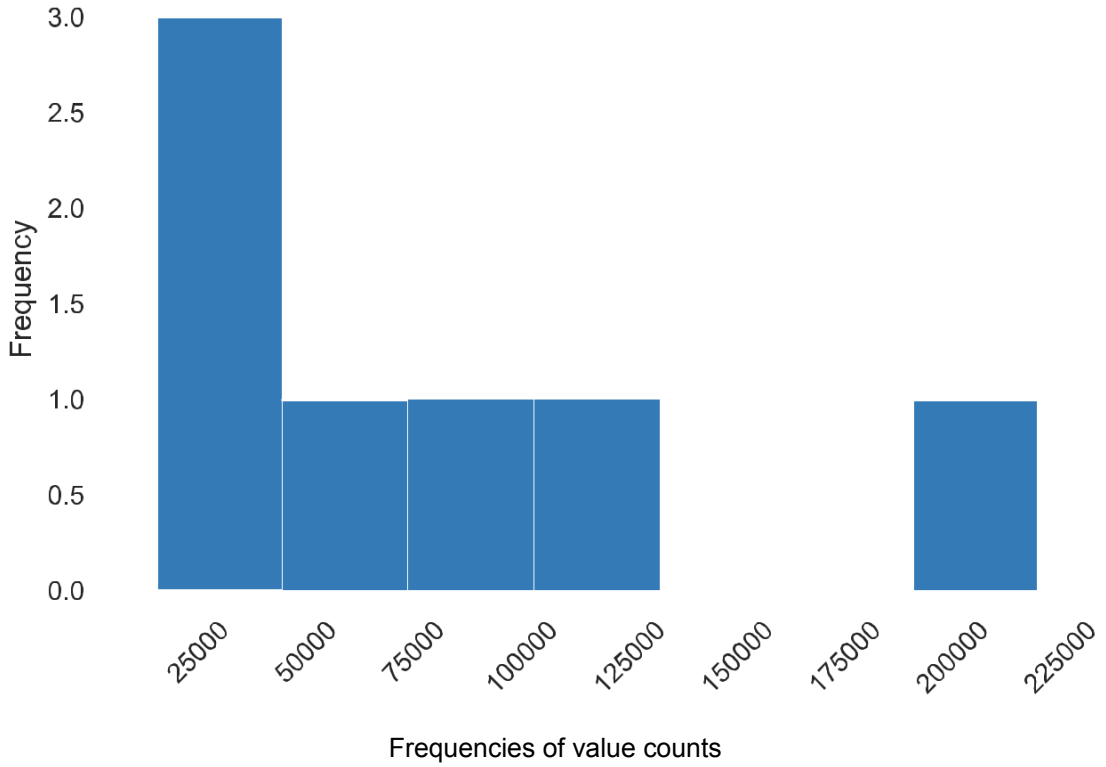
Distinct	7
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.2 MiB

Frequencies

Value	Count	Frequency (%)
26-35	219587	39.9%
36-45	110013	20.0%
18-25	99660	18.1%
46-50	45701	8.3%
51-55	38501	7.0%
55+	21504	3.9%
0-17	15102	2.7%

Common Values

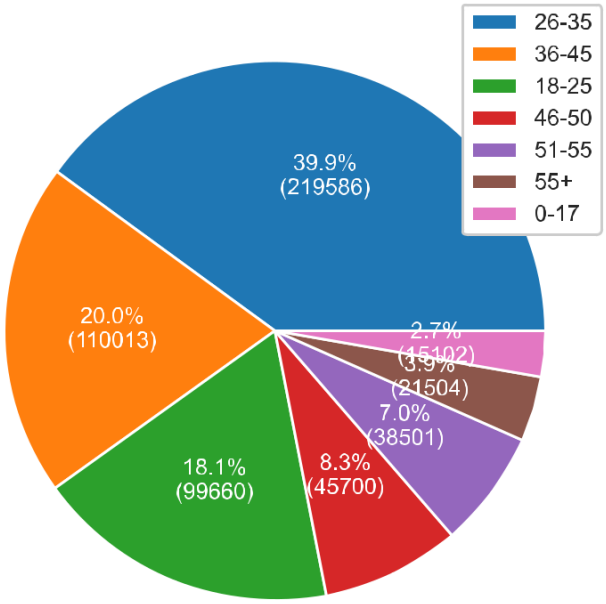
Overview



Unique

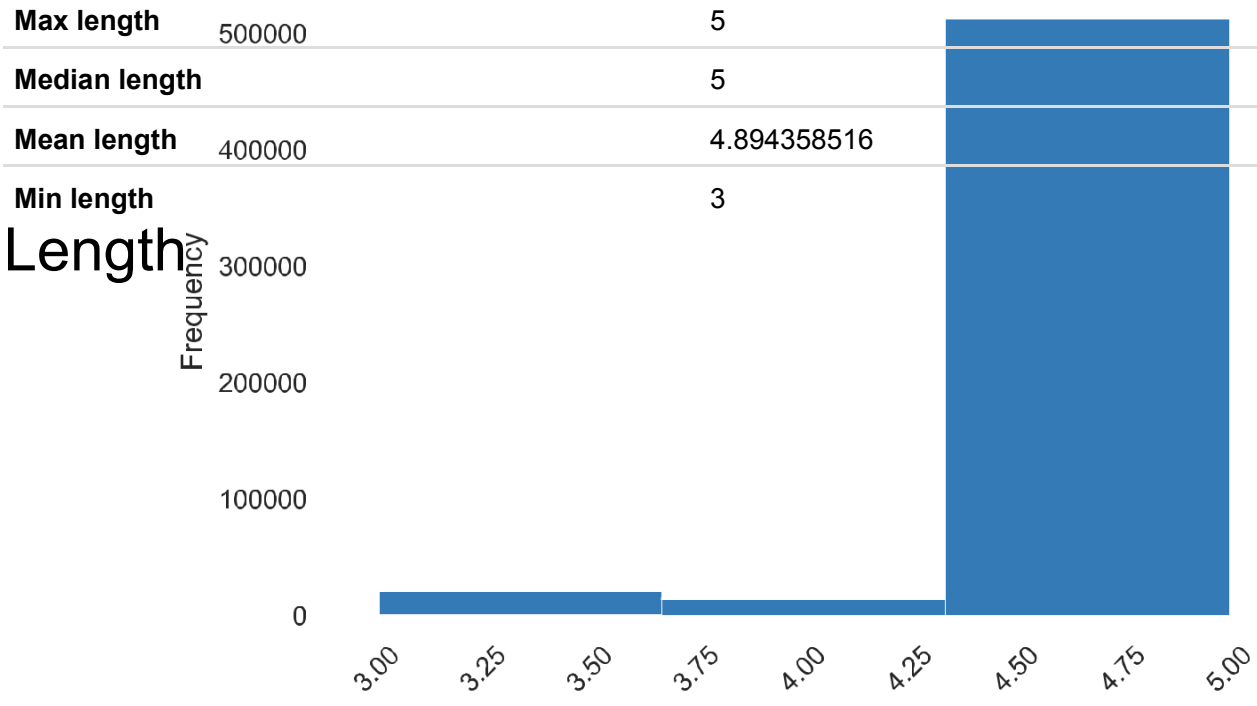
Unique	0	**?**
Unique (%)	0.0%	

Chart



Length

Max length	5
Median length	5
Mean length	4.894358516
Min length	3



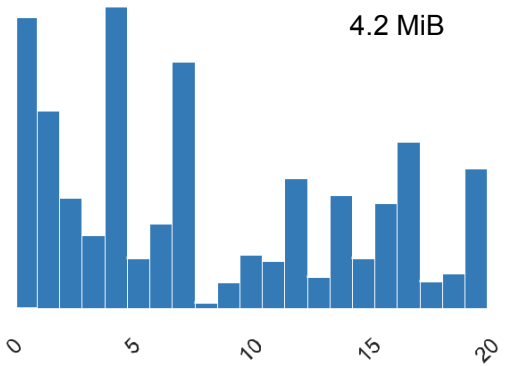
Histogram of lengths of the category

Occupation

Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct	21
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	8.07670688
Minimum	0
Maximum	20
Zeros	69638
Zeros (%)	12.7%
Memory size	4.2 MiB



Statistics

Quantile statistics

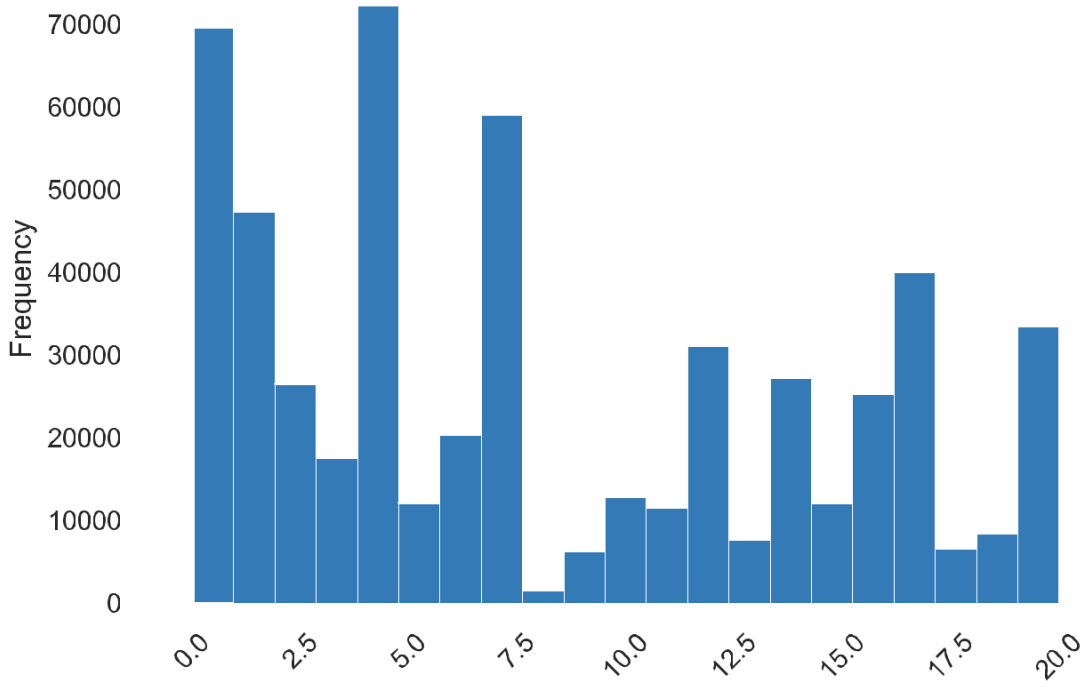
Minimum	0
5-th percentile	0
Q1	2
median	7
Q3	14
95-th percentile	20
Maximum	20
Range	20
Interquartile range (IQR)	12

Descriptive statistics

Standard deviation	6.522660487
Coefficient of variation (CV)	0.8075891059

Kurtosis	-1.216113649
Mean	8.07670688
Median Absolute Deviation (MAD)	6
Skewness	0.4001401099
Sum	4442738
Variance	42.54509983
Monotocity	Not monotonic

Histogram



Histogram with fixed size bins (bins=21)

Common Values

Value	Count	Frequency (%)
4	72308	13.1%
0	69638	12.7%
7	59133	10.8%
1	47426	8.6%

17	40043	7.3%

Value	Count	Frequency (%)
20	33562	6.1%
12	31179	5.7%
14	27309	5.0%
2	26588	4.8%
16	25371	4.6%
Other values (11)	117511	21.4%

Extreme Values

Minimum 5 values

Value	Count	Frequency (%)
0	69638	12.7%
1	47426	8.6%
2	26588	4.8%
3	17650	3.2%
4	72308	13.1%

Maximum 5 values

Value	Count	Frequency (%)
20	33562	6.1%
19	8461	1.5%
18	6622	1.2%
17	40043	7.3%
16	25371	4.6%

City_Category

Categorical

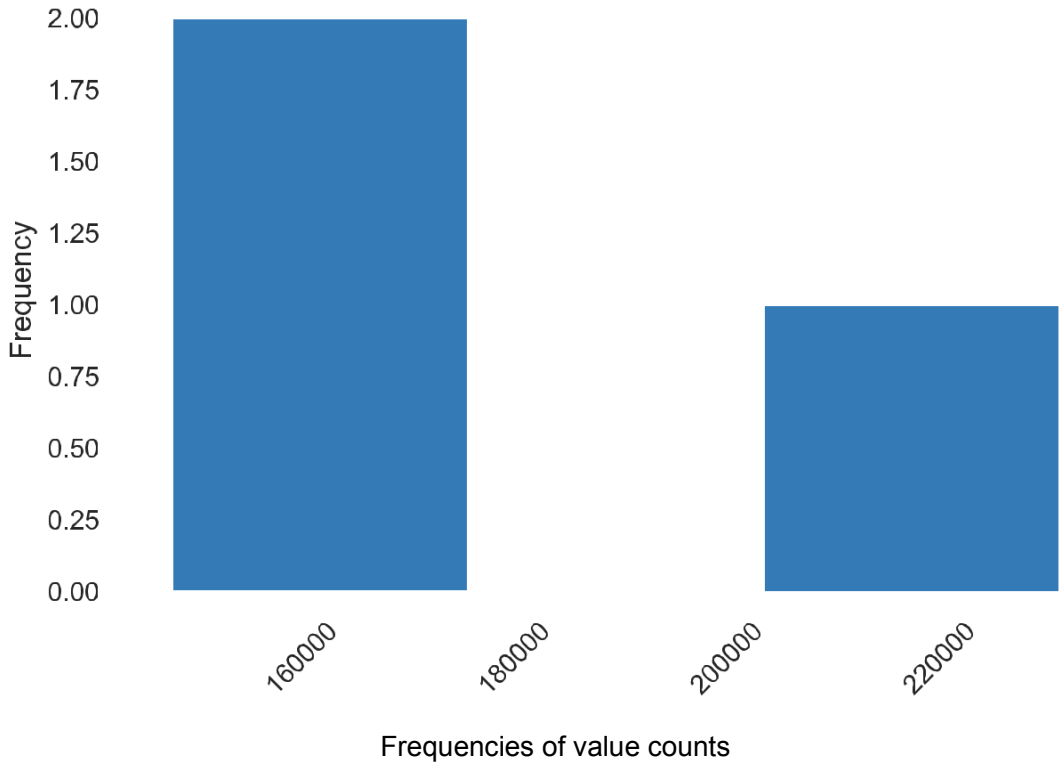
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.2 MiB

Frequencies

Common Values

Value	Count	Frequency (%)
B	231173	42.0%
C	171175	31.1%
A	147720	26.9%

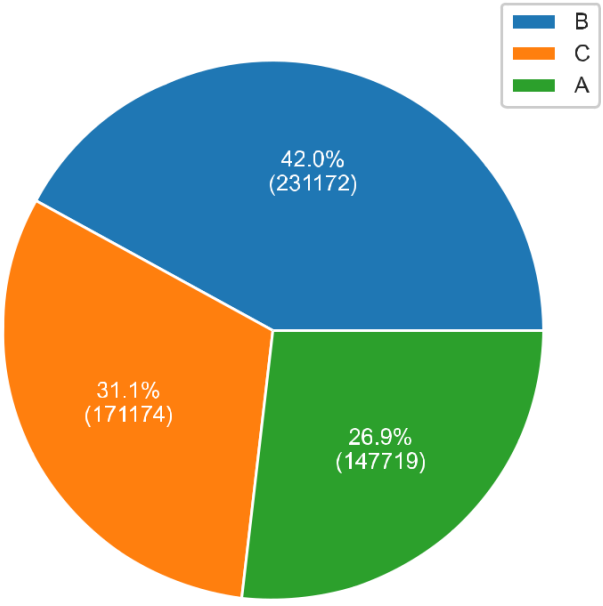
Overview



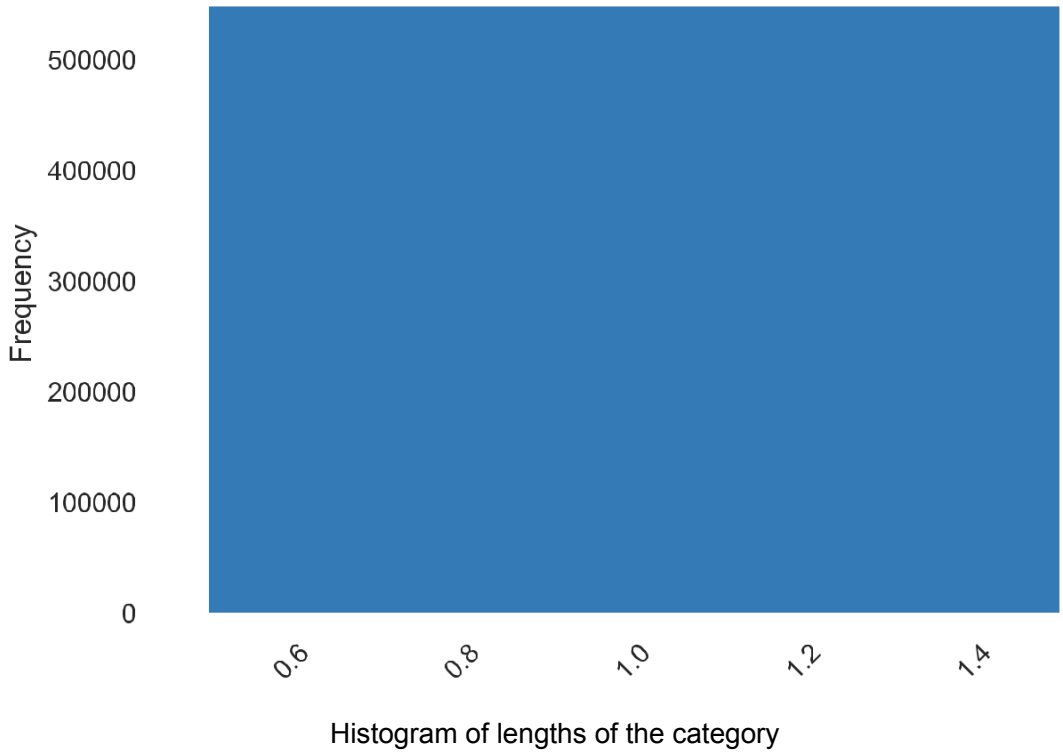
Unique

Unique	0	**?**
Unique (%)	0.0%	

Chart



Length



Histogram of lengths of the category

Length

Max length	1
Median length	1
Mean length	1
Min length	1

Stay_In_Current_City_Years

Categorical

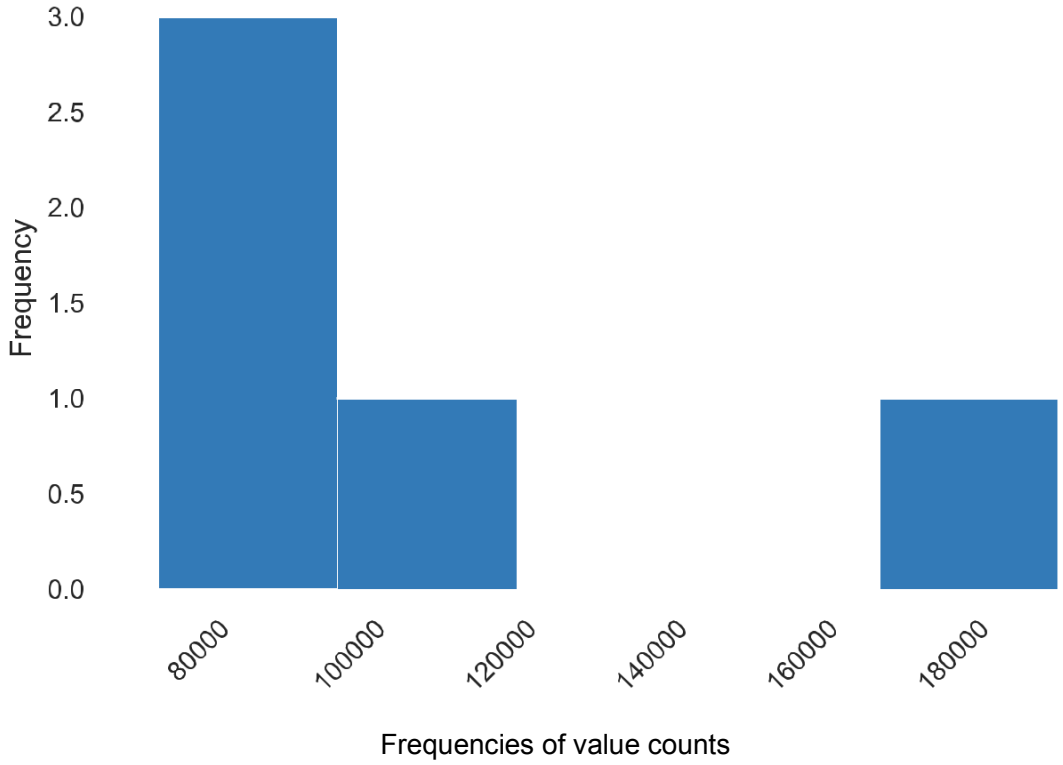
Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.2 MiB

Frequencies

Common Values

Value	Count	Frequency (%)
1	193821	35.2%
2	101838	18.5%
3	95285	17.3%
4+	84726	15.4%
0	74398	13.5%

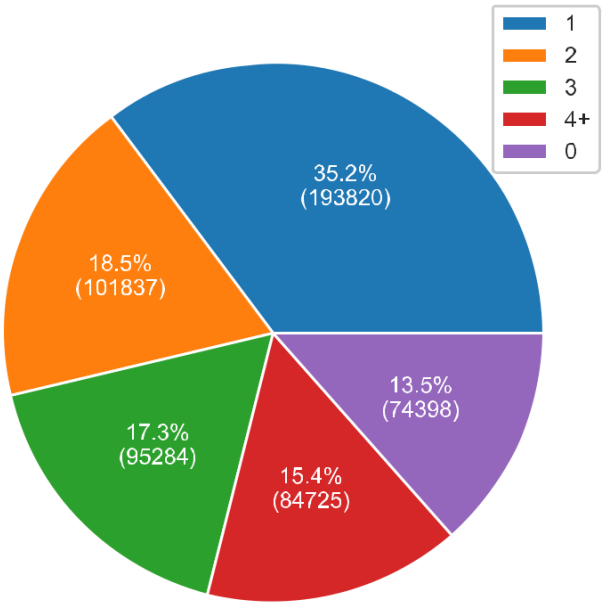
Overview



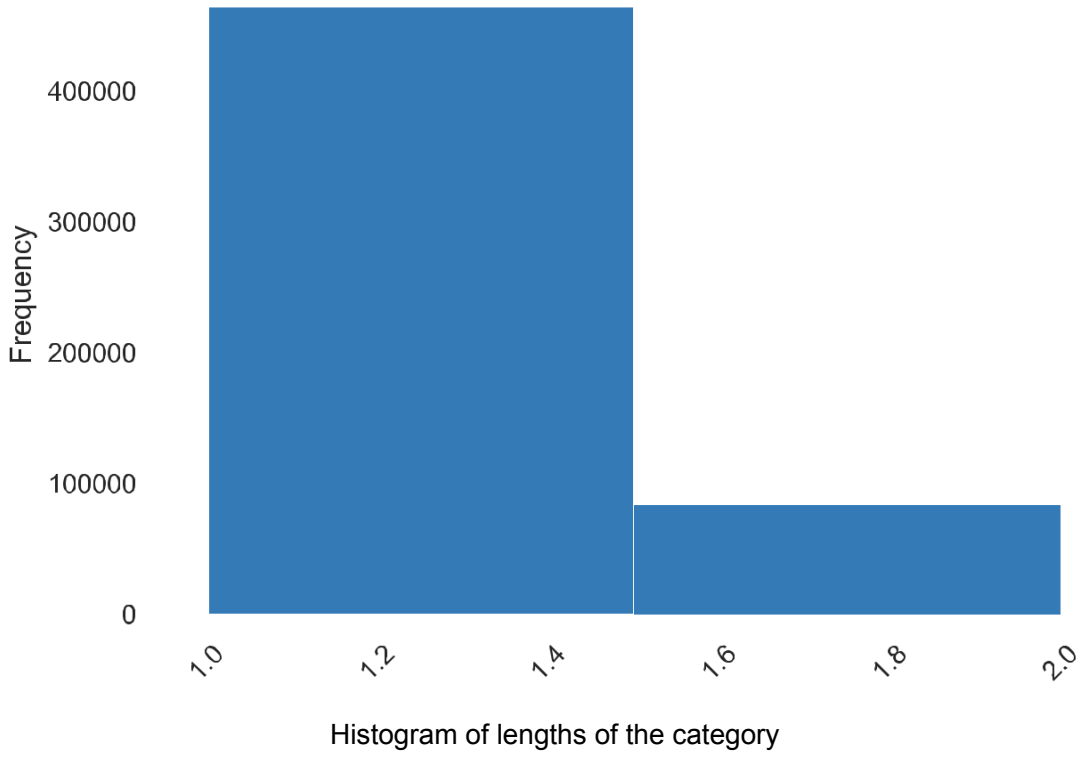
Unique

Unique	0	**?**
Unique (%)	0.0%	

Chart



Length



Length

Max length	2
Median length	1
Mean length	1.154028229
Min length	1

Marital_Status

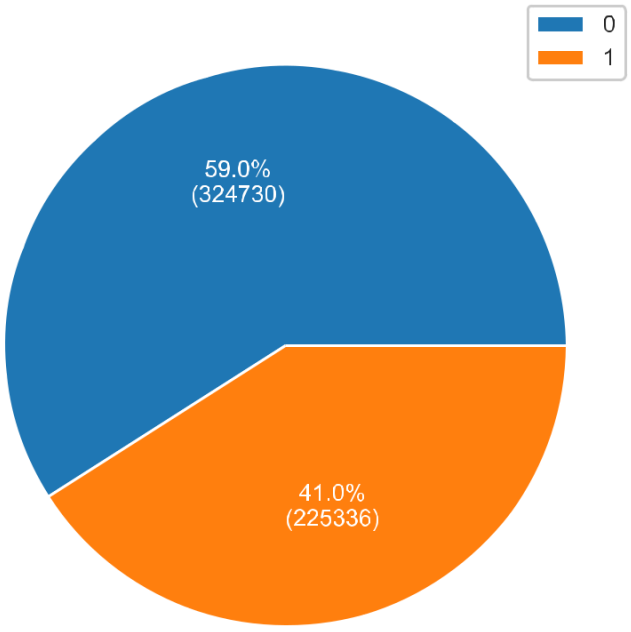
Boolean

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	4.2 MiB

Common Values

Value	Count	Frequency (%)
0	324731	59.0%
1	225337	41.0%

Chart



Product_Category_1

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	20
Distinct (%)	< 0.1%

Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

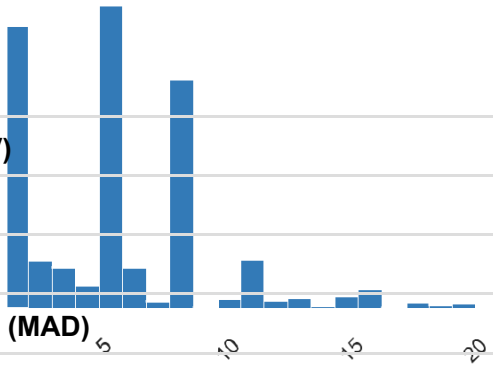
Mean	5.404270018
Minimum	1
Maximum	20
Zeros	0
Zeros (%)	0.0%
Memory size	4.2 MiB

Statistics

Quantile statistics

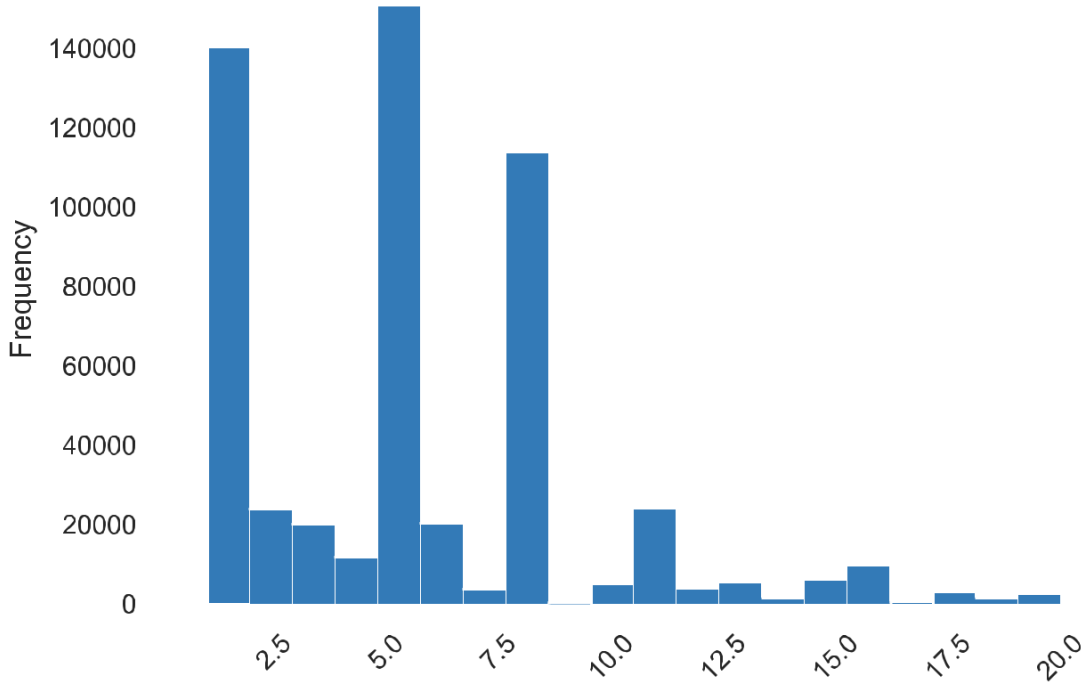
Minimum	1
5-th percentile	1
Q1	1
median	5
Q3	8
95-th percentile	13
Maximum	20
Range	19
Interquartile range (IQR)	7

Descriptive statistics



Standard deviation	3.936211369
Coefficient of variation (CV)	0.7283520913
Kurtosis	1.234756972
Mean	5.404270018
Median Absolute Deviation (MAD)	3
Skewness	1.025734934
Sum	2972716
Variance	15.49375994
Monotocity	Not monotonic

Histogram



Histogram with fixed size bins (bins=20)

Common Values

Value	Count	Frequency (%)
5	150933	27.4%
1	140378	25.5%

Value	Count	Frequency (%)
8	113925	20.7%
11	24287	4.4%
2	23864	4.3%
6	20466	3.7%
3	20213	3.7%
4	11753	2.1%
16	9828	1.8%
15	6290	1.1%
Other values (10)	28131	5.1%

Extreme Values

Minimum 5 Values

Value	Count	Frequency (%)
1	140378	25.5%
2	23864	4.3%
3	20213	3.7%
4	11753	2.1%
5	150933	27.4%

Maximum 5 Values

Value	Count	Frequency (%)
20	2550	0.5%
19	1603	0.3%
18	3125	0.6%
17	578	0.1%
16	9828	1.8%

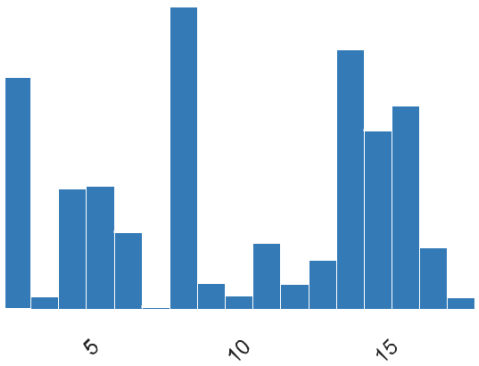
Product_Category_2

Real number ($\mathbb{R}_{\geq 0}$)

MISSING

Distinct	17
Distinct (%)	< 0.1%
Missing	173638
Missing (%)	31.6%
Infinite	0
Infinite (%)	0.0%

Mean	9.842329251
Minimum	2
Maximum	18
Zeros	0
Zeros (%)	0.0%
Memory size	4.2 MiB



Statistics

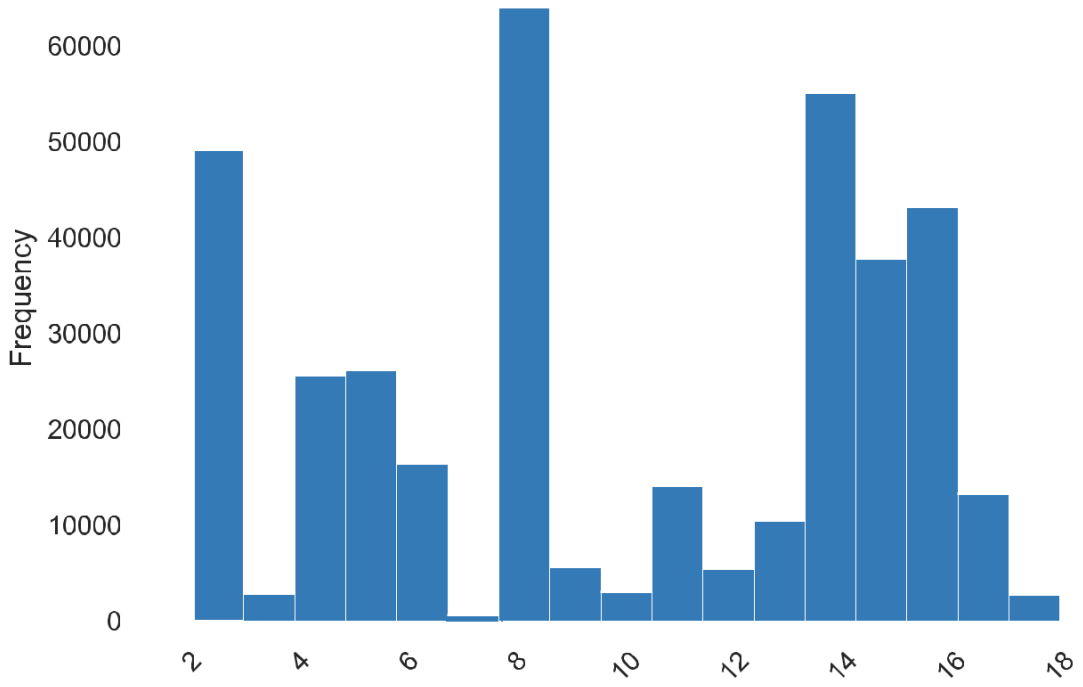
Quantile statistics

Minimum	2
5-th percentile	2
Q1	5
median	9
Q3	15
95-th percentile	16
Maximum	18
Range	16
Interquartile range (IQR)	10

Descriptive statistics

Standard deviation	5.086589649
Coefficient of variation (CV)	0.5168075075
Kurtosis	-1.432266899
Mean	9.842329251
Median Absolute Deviation (MAD)	5
Skewness	-0.1627577144
Sum	3704948
Variance	25.87339425
Monotocity	Not monotonic

Histogram



Histogram with fixed size bins (bins=17)

Common Values

Value	Count	Frequency (%)
8	64088	11.7%
14	55108	10.0%
2	49217	8.9%
16	43255	7.9%
15	37855	6.9%
5	26235	4.8%
4	25677	4.7%
6	16466	3.0%
11	14134	2.6%
17	13320	2.4%
Other values (7)	31075	5.6%
(Missing)	173638	31.6%

Extreme Values

Minimum 5 Values

Value	Count	Frequency (%)
2	49217	8.9%
3	2884	0.5%
4	25677	4.7%
5	26235	4.8%
6	16466	3.0%

Maximum 5 Values

Value	Count	Frequency (%)
18	2770	0.5%
17	13320	2.4%
16	43255	7.9%
15	37855	6.9%
14	55108	10.0%

Product_Category_3

Real number ($\mathbb{R}_{\geq 0}$)

MISSING

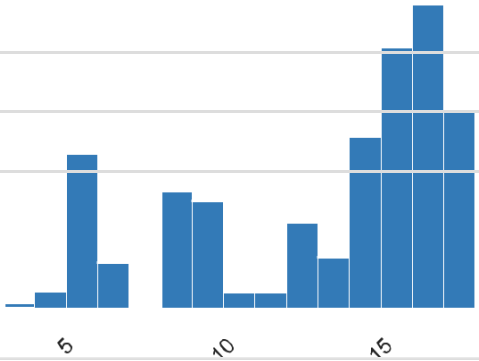
Distinct	15
Distinct (%)	< 0.1%
Missing	383247
Missing (%)	69.7%
Infinite	0
Infinite (%)	0.0%

Mean	12.66824321
Minimum	3
Maximum	18
Zeros	0
Zeros (%)	0.0%
Memory size	4.2 MiB

Statistics

Quantile statistics

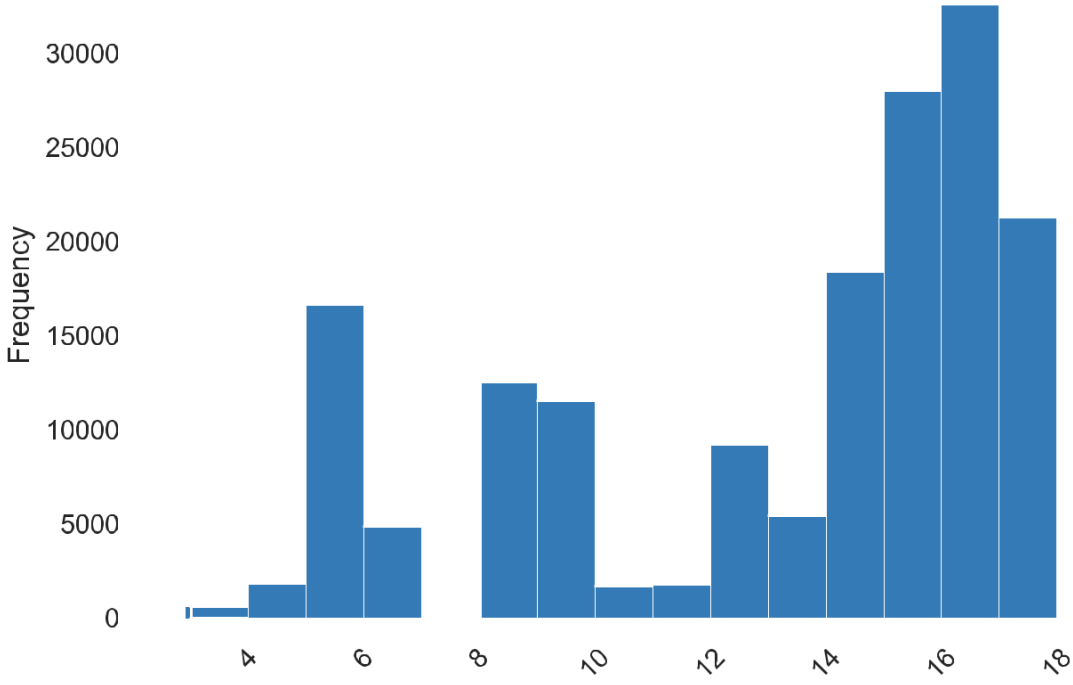
Minimum	3
5-th percentile	5
Q1	9
median	14
Q3	16
95-th percentile	17
Maximum	18
Range	15
Interquartile range (IQR)	7



Descriptive statistics

Standard deviation	4.125337632
Coefficient of variation (CV)	0.325644019
Kurtosis	-0.8080661151
Mean	12.66824321
Median Absolute Deviation (MAD)	2
Skewness	-0.7654458894
Sum	2113329
Variance	17.01841057
Monotocity	Not monotonic

Histogram



Histogram with fixed size bins (bins=15)

Common Values

Value	Count	Frequency (%)
16	32636	5.9%
15	28013	5.1%
14	18428	3.4%
17	16702	3.0%
Value	**Count**	**Frequency (%)**
5	16658	3.0%
8	12562	2.3%
9	11579	2.1%
12	9246	1.7%
13	5459	1.0%
6	4890	0.9%
Other values (5)	10648	1.9%
(Missing)	383247	69.7%

Extreme Values

Minimum 5 Values

Maximum 5 Values

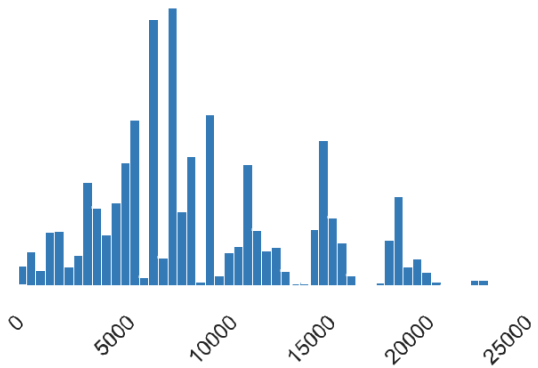
Value	Count	Frequency (%)
18	4829	0.8%
17	16702	3.0%
16	32636	5.9%
15	28013	5.1%
8	12562	2.3%

Purchase

Distinct	18105
Distinct (%)	3.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	9263.968713
Minimum	12
Maximum	23961
Zeros	0
Zeros (%)	0.0%
Memory size	4.2 MiB

Real number ($\mathbb{R}_{\geq 0}$)



Statistics

Quantile statistics

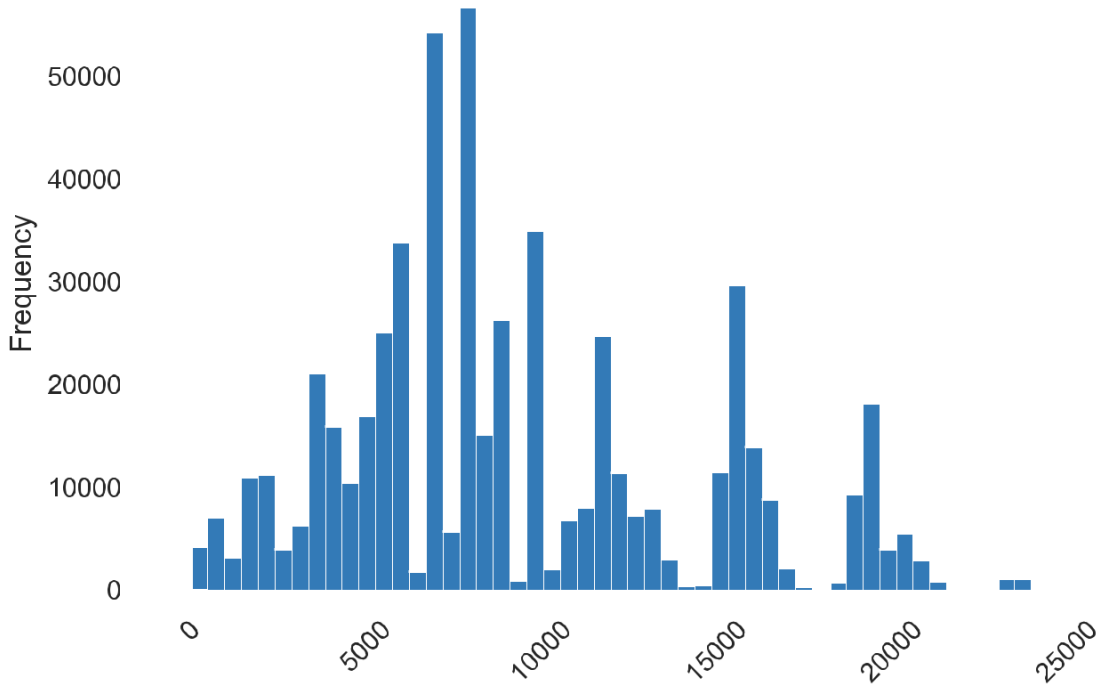
Minimum	12
5-th percentile	1984
Q1	5823

median	8047
Q3	12054
95-th percentile	19336
Maximum	23961
Range	23949
Interquartile range (IQR)	6231

Descriptive statistics

Standard deviation	5023.065394
Coefficient of variation (CV)	0.5422152805
Kurtosis	-0.3383775656
Mean	9263.968713
Median Absolute Deviation (MAD)	2871
Skewness	0.6001400037
Sum	5095812742
Variance	25231185.95
Monotocity	Not monotonic

Histogram



Histogram with fixed size bins (bins=50)

Common Values

Value	Count	Frequency (%)
7011	191	< 0.1%
7193	188	< 0.1%
6855	187	< 0.1%
6891	184	< 0.1%
6960	183	< 0.1%
7012	183	< 0.1%
6879	182	< 0.1%
7166	182	< 0.1%
7027	182	< 0.1%
7165	180	< 0.1%
Other values (18095)	548226	99.7%

Extreme Values

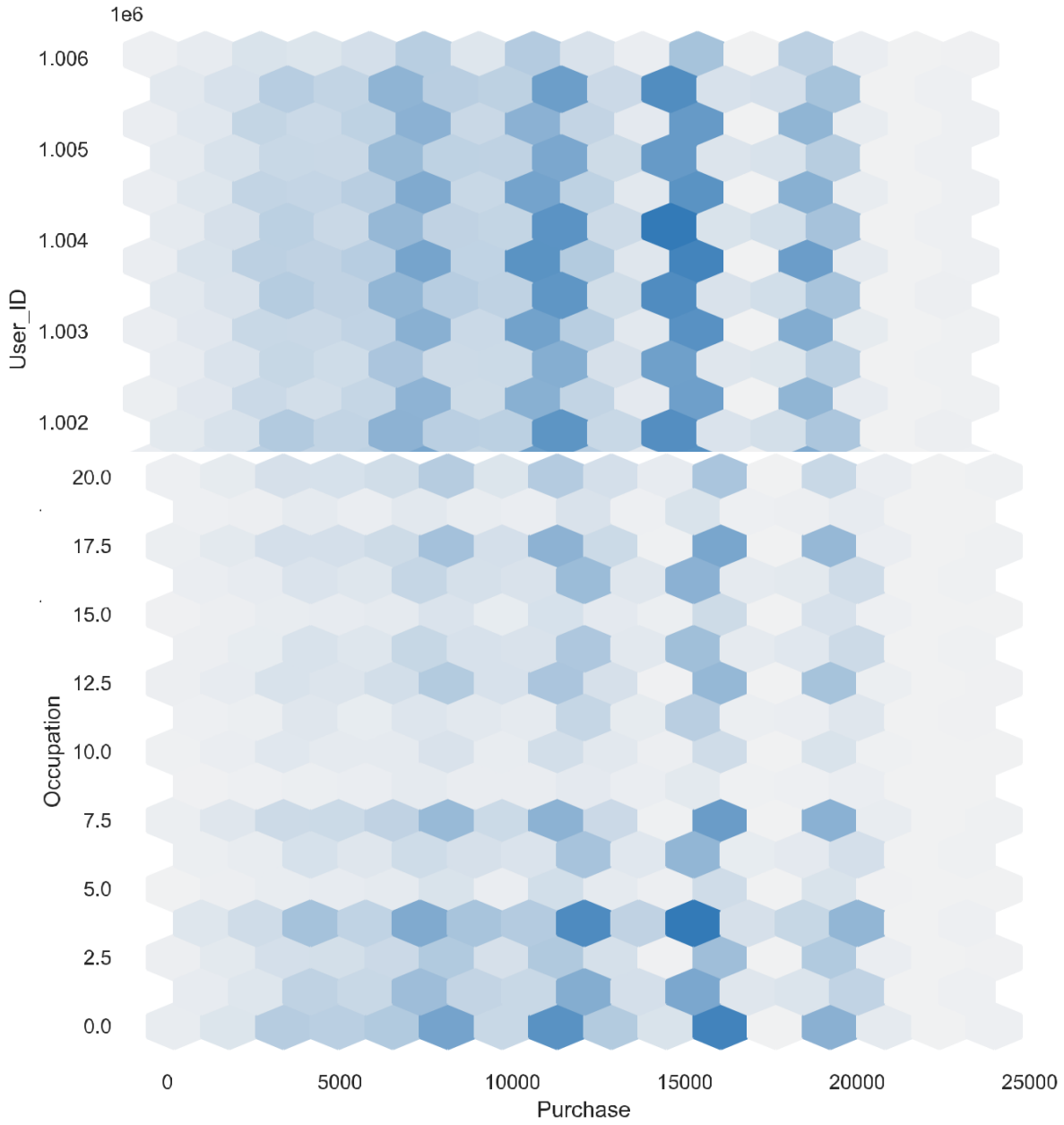
Minimum 5 Values

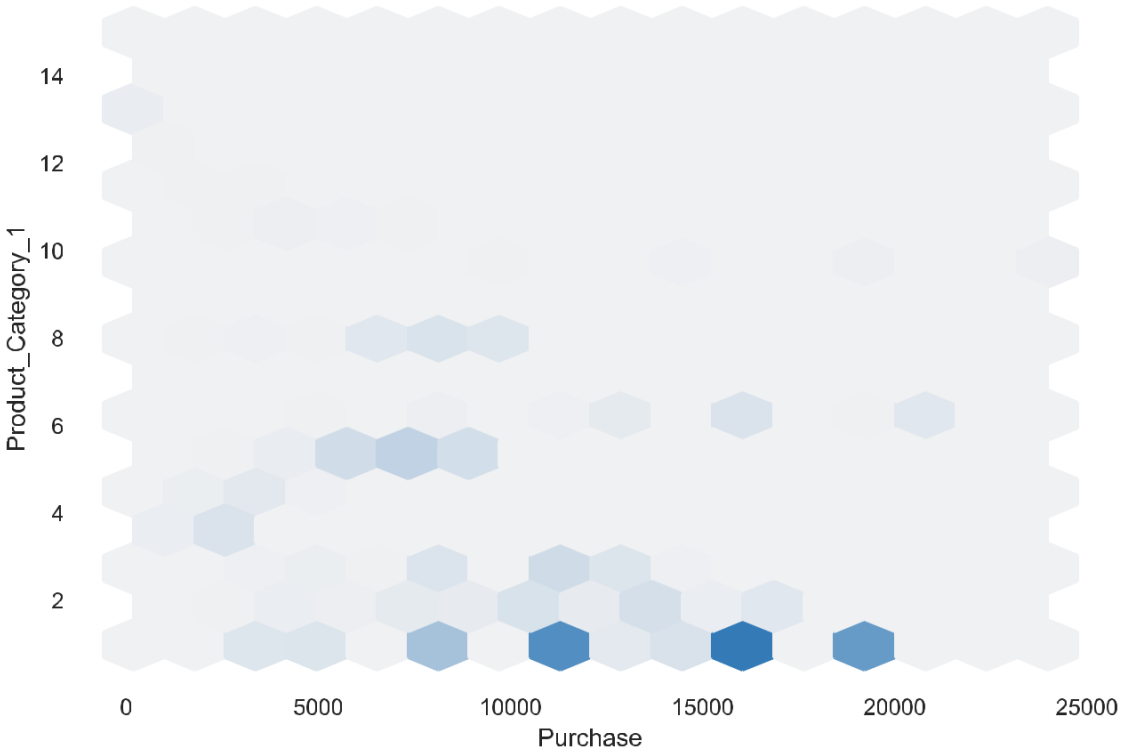
Value	Count	Frequency (%)
12	101	< 0.1%
13	106	< 0.1%
14	95	< 0.1%
24	118	< 0.1%
25	113	< 0.1%

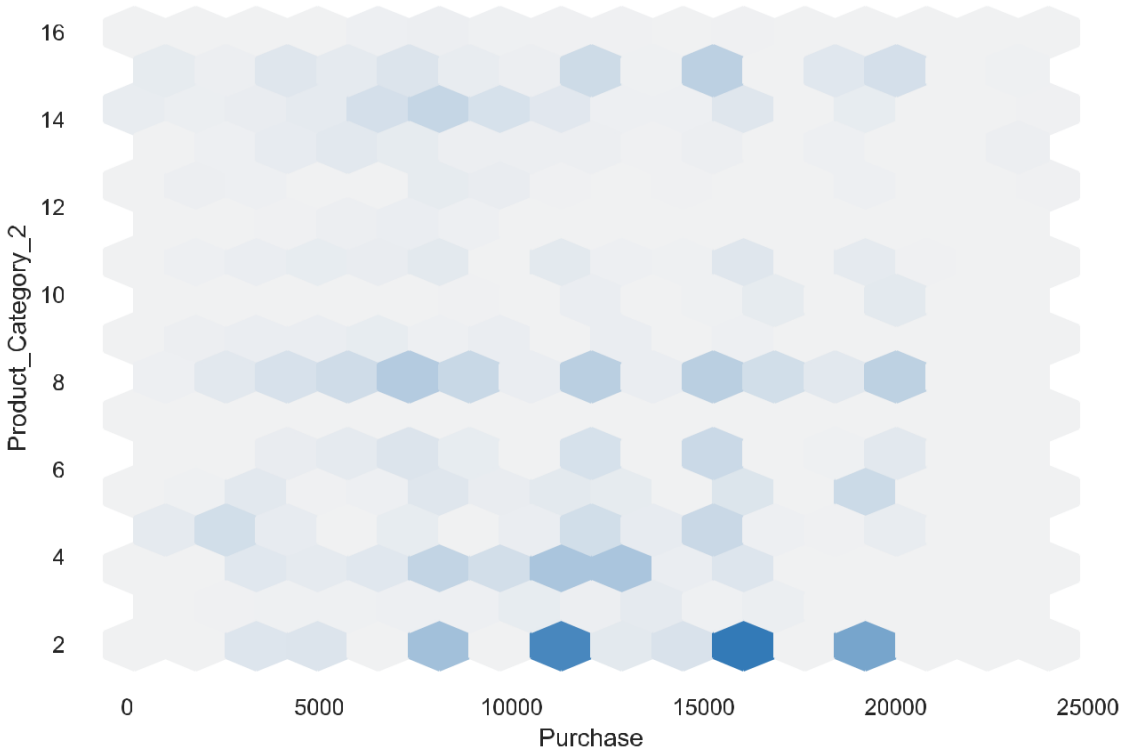
Maximum 5 Values

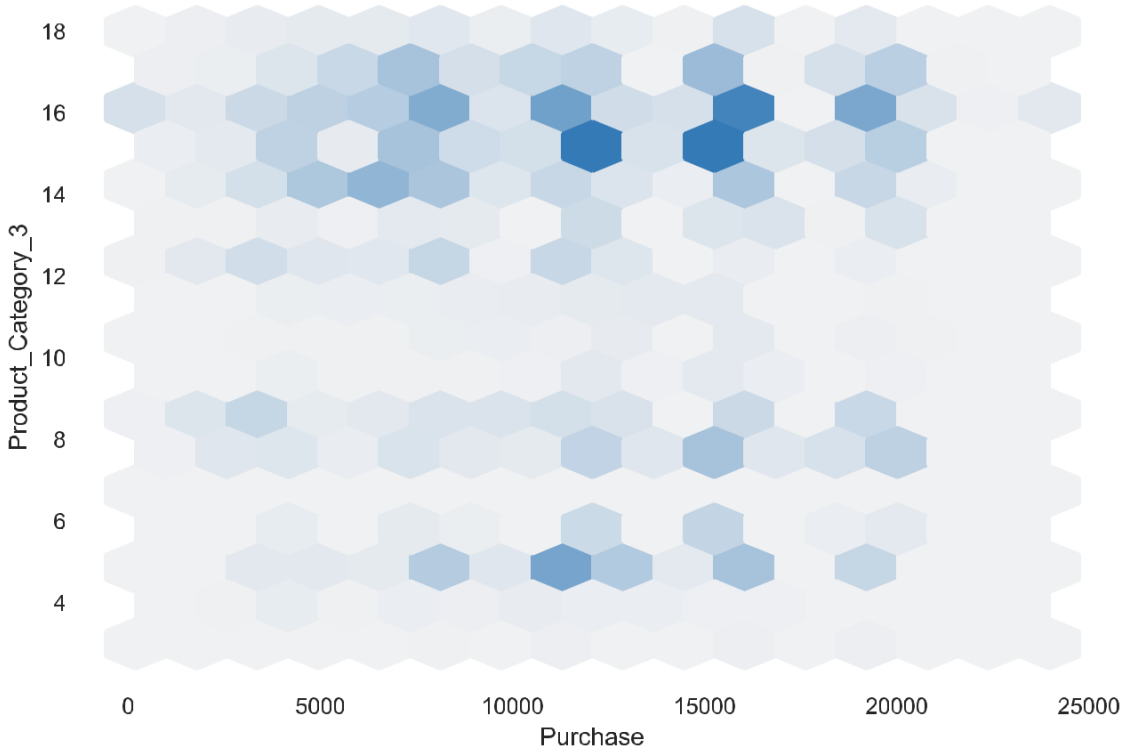
Value	Count	Frequency (%)
23961	3	< 0.1%
23960	4	< 0.1%
23959	2	< 0.1%
23958	4	< 0.1%
23956	1	< 0.1%

Interactions

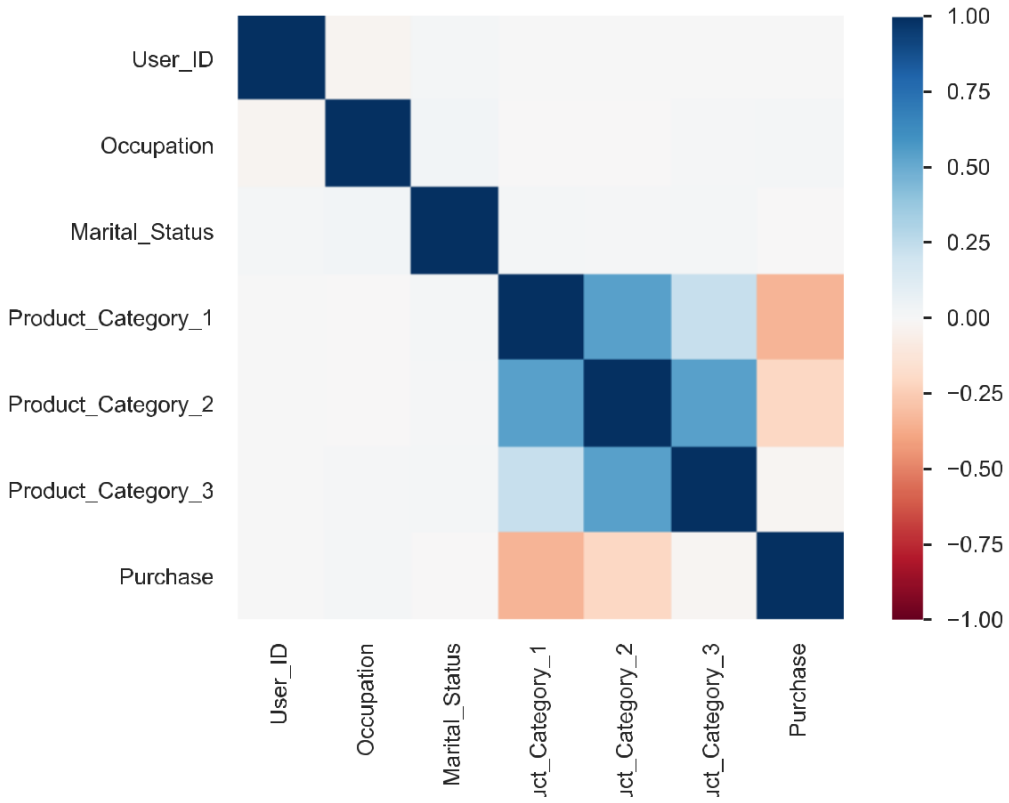



Product_Category_1
14
12
10
8
6
4
2
0
5000
10000
15000
20000
25000
Purchase


Product_Category_2
16
14
12
10
8
6
4
2
0
5000
10000
15000
20000
25000
Purchase


18
16
14
12
10
8
6
4
Product_Category_3
0
5000
10000
15000
20000
25000
Purchase

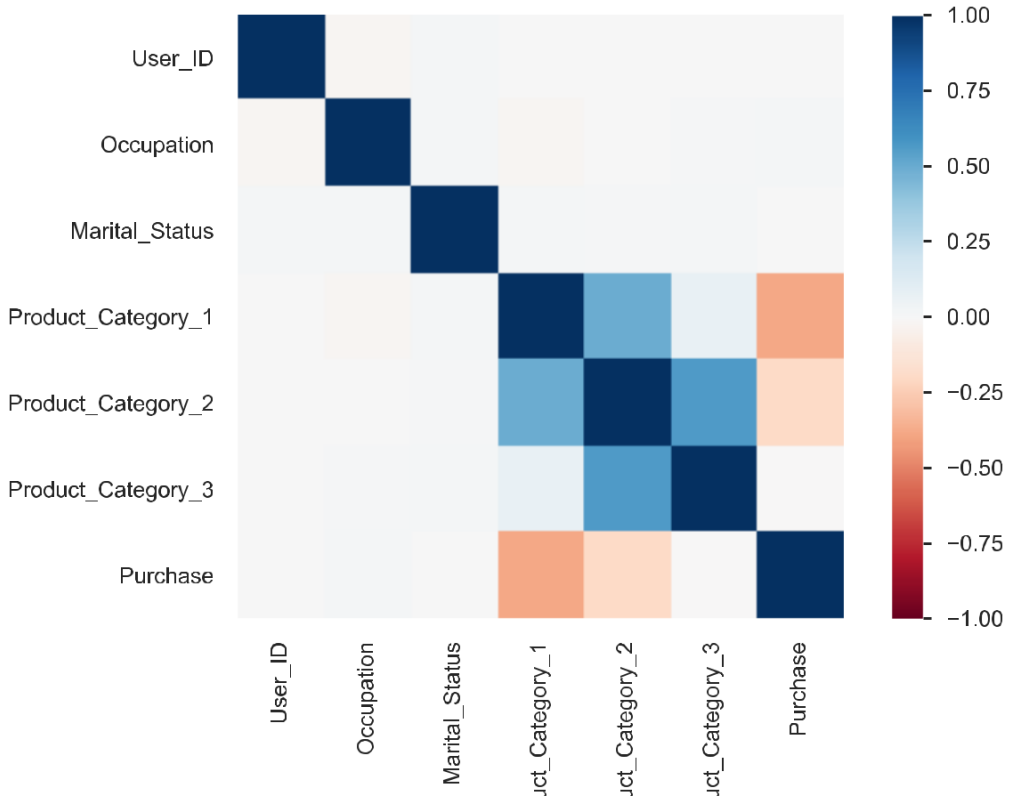
Correlations



Pearson's r

The Pearson's correlation coefficient (*r*) is a measure of linear correlation between two variables. It's value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, *r* is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect *r*.

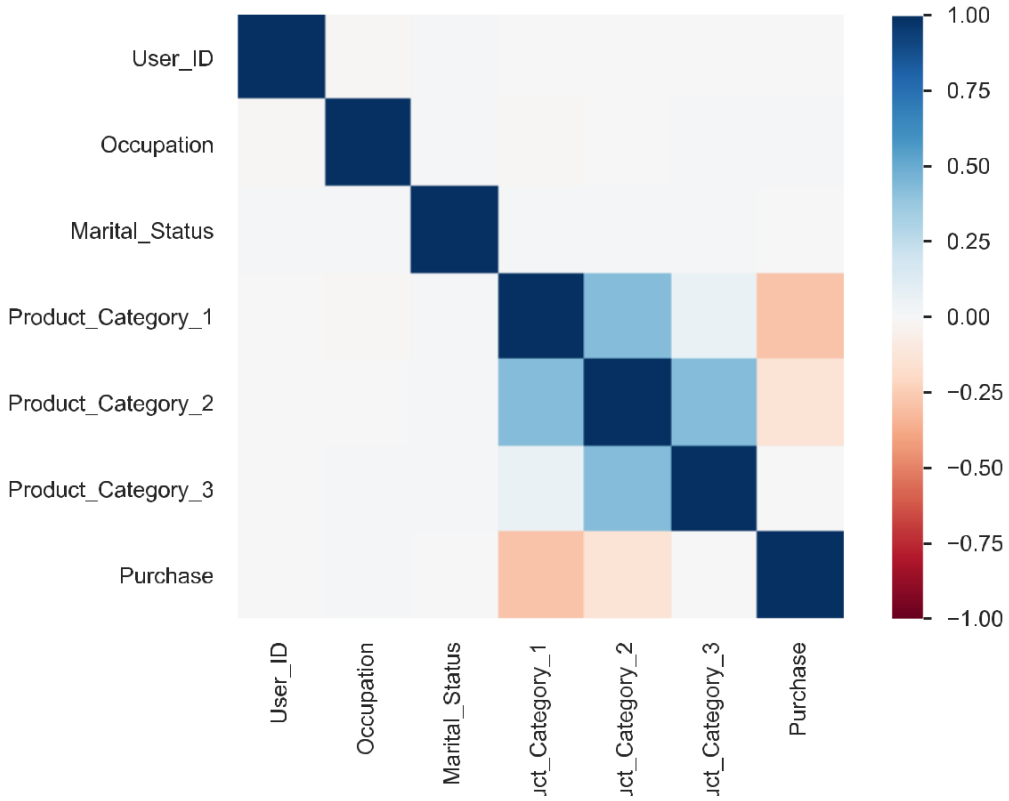
To calculate *r* for two variables *X* and *Y*, one divides the covariance of *X* and *Y* by the product of their standard deviations.



Spearman's ρ

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's *r*. It's value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

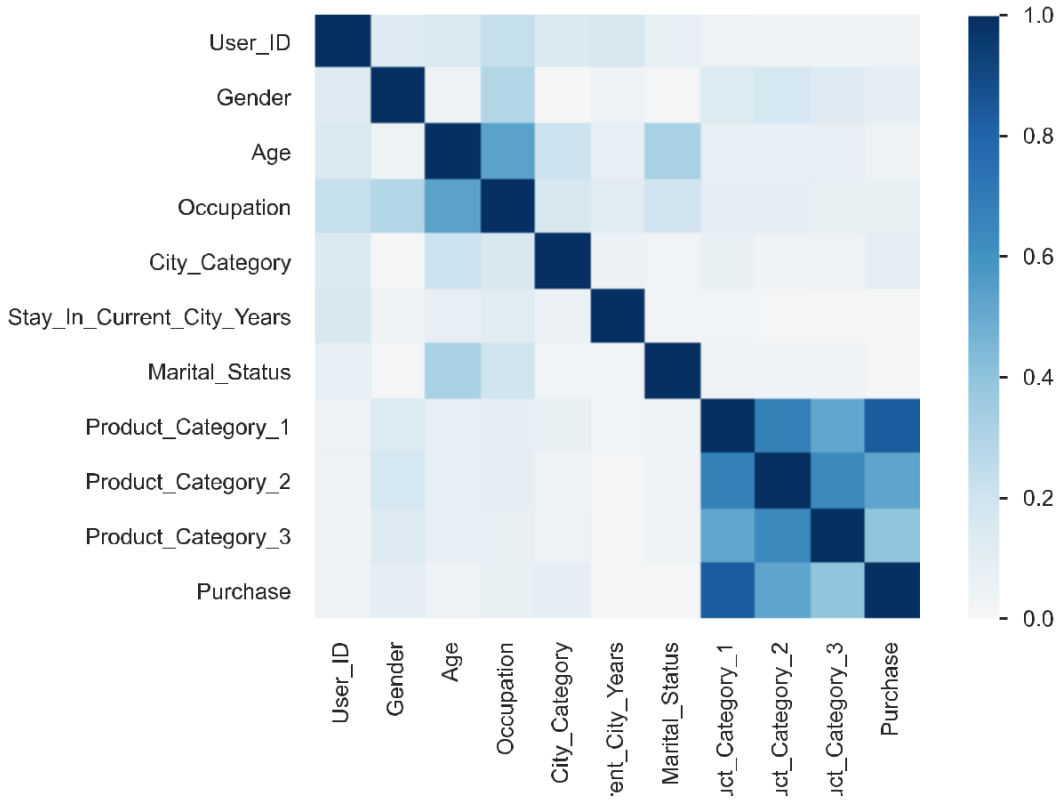
To calculate ρ for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.



Kendall's τ

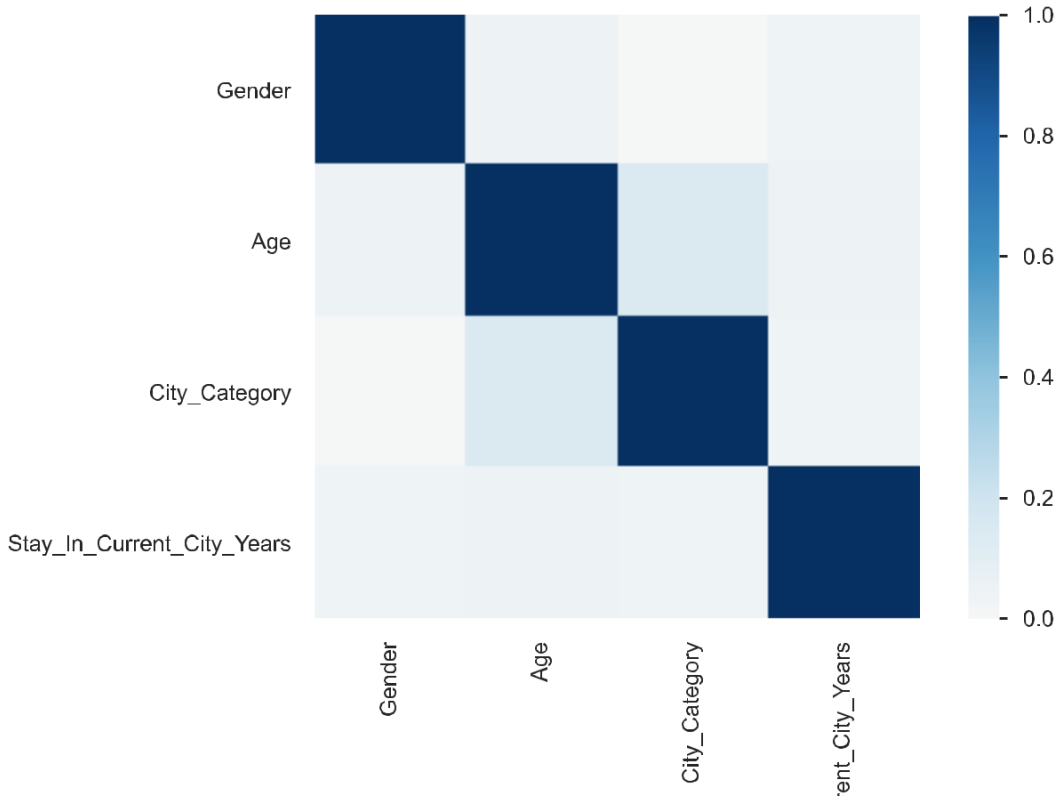
Similarly to Spearman's rank correlation coefficient, the Kendall rank correlation coefficient (τ) measures ordinal association between two variables. It's value lies between -1 and +1, -1 indicating total negative correlation, 0 indicating no correlation and 1 indicating total positive correlation.

To calculate τ for two variables X and Y, one determines the number of concordant and discordant pairs of observations. τ is given by the number of concordant pairs minus the discordant pairs divided by the total number of pairs.



Phik (φk)

Phik (φk) is a new and practical correlation coefficient that works consistently between categorical, ordinal and interval variables, captures non-linear dependency and reverts to the Pearson correlation coefficient in case of a bivariate normal input distribution. There is extensive documentation available [here (https://phik.readthedocs.io/en/latest/index.html)](https://phik.readthedocs.io/en/latest/index.html).

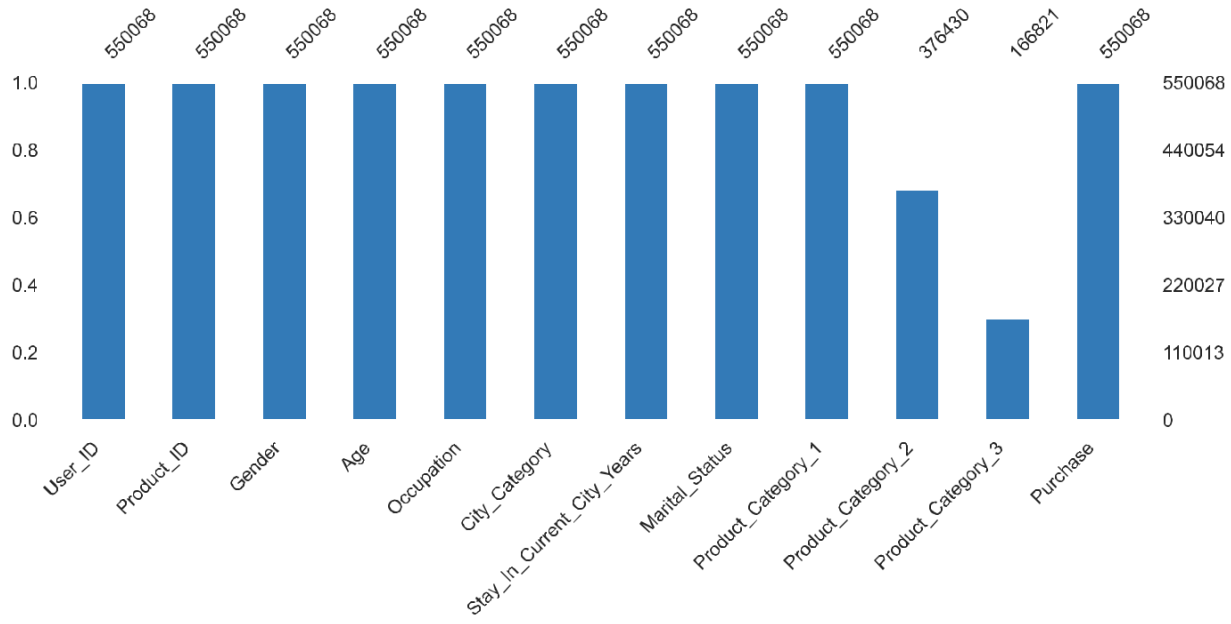


Cramér's V (φc)

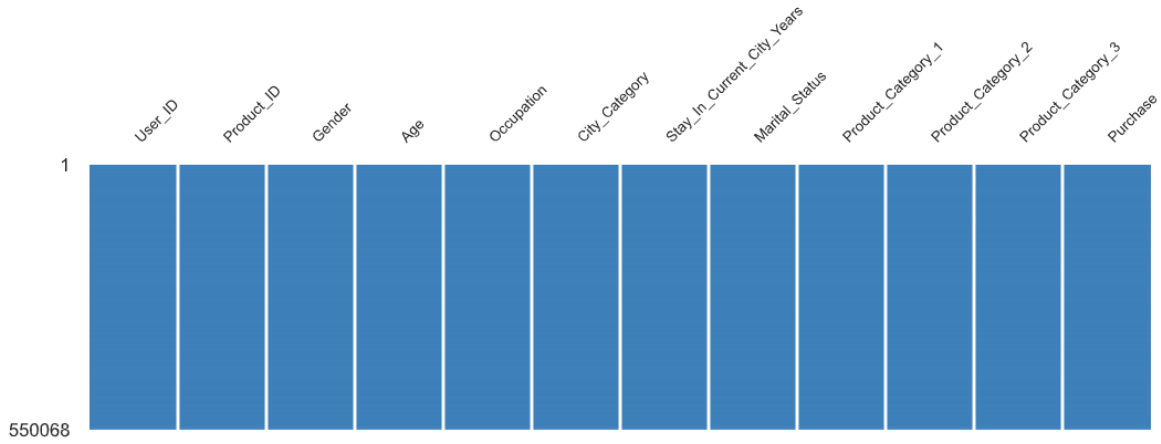
Cramér's V is an association measure for nominal random variables. The coefficient ranges from 0 to 1, with 0 indicating independence and 1 indicating perfect association. The empirical estimators used for Cramér's V have been proved to be biased, even for large samples. We use a bias-corrected measure that has been proposed by Bergsma in 2013 that can be found here (http://stats.lse.ac.uk/bergsma/pdf/cramerV3.pdf).

Missing values

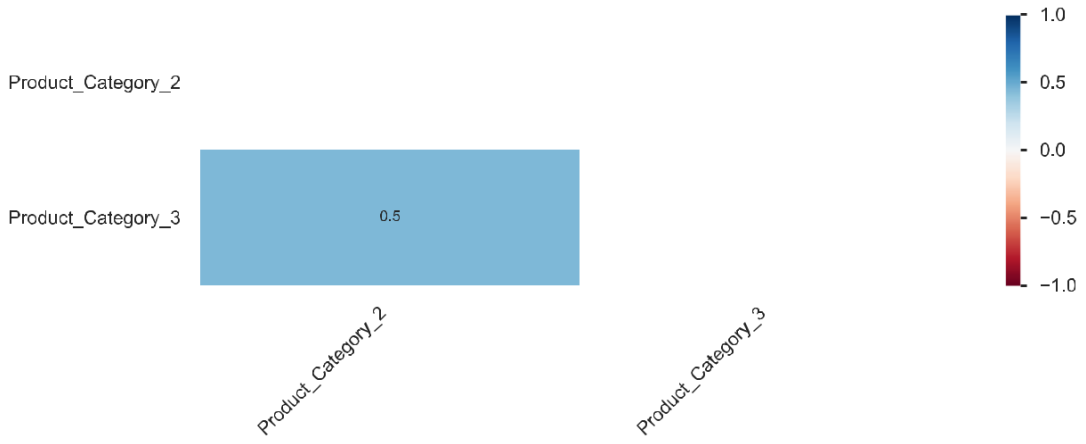
Count



Matrix



Heatmap



Dendrogram