KOBOLD.AI

Customer Data Quality Analysis

Data quality highly influences the results of any further processing, be it data logistics, analytics or artificial intelligence. Here we analyse the data set with regard to six dimensions of data quality and give according recommendations for next steps.

Meta-Data

Algorithm:	Customer Data Quality
Start:	07-Sep-2022 (19:40:59.614629)
Filename:	pak_ecom.csv
Dimensions:	(50000, 26)

Input data overview

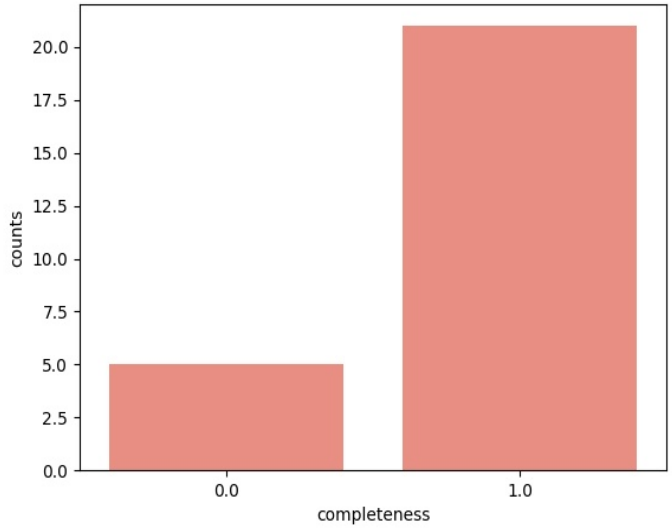
Attribute	Count	Mean	Range	Mode	Unique
item_id	50000	241987.69	211131 - 273201	211131	50000
status	50000	-	-	complete	8
created_at	50000	-	-	9/30/2016	126
sku	49995	-	-	kcc_krone deal	8747
price	50000	4520.59	0.0 - 300000.0	399.0	1770
qty_ordered	50000	1.15	1 - 130	1	25
grand_total	50000	5931.0	0.0 - 1280473.0	360.0	3462
increment_id	50000	100168528.99	100147443 - 100189643	100154653	42192

KOBOLD.AI

Attribute	Count	Mean	Range	Mode	Unique
category_name_1	50000	-	-	Soghaat	16
sales_commission_code	50000	-	-	\N	2033
discount_amount	50000	112.46	0.0 - 33750.0	0.0	401
payment_method	50000	-	-	cod	11
Working Date	50000	-	-	9/30/2016	126
BI Status	50000	-	-	Net	4
MV	50000	-	-	399	2089
Year	50000	2016.0	2016 - 2016	2016	1
Month	50000	8.72	7 - 11	9	5
Customer Since	50000	-	-	2016-7	5
M-Y	50000	-	-	9-2016	5
FY	50000	-	-	FY17	1
Customer ID	50000	4416.62	1 - 13077	35	13077
Unnamed: 21	0	nan	nan - nan	-	0
Unnamed: 22	0	nan	nan - nan	-	0
Unnamed: 23	0	nan	nan - nan	-	0
Unnamed: 24	0	nan	nan - nan	-	0
Unnamed: 25	0	nan	nan - nan	-	0

KOBOLD.AI

Completeness



Completeness is the value of how many empty cells an attribute has. A low completeness can have many reasons, but you should be aware that the data does have missing data. Usually, the aim is to provide 100% completeness by utilising NULL-values such as "0" or "NULL", to differentiate between missing values and values that define empty cells.

In this graph we defined how many columns exhibit what percentage of completeness. If you have all at 100%, your data is well filled. If any attributes are below 100%, they should be analysed why data is missing and whether it can be prevented. If there are columns with 0% completeness, there are empty columns in your data set.

Attribute	Completeness
item_id	1.0
status	1.0
created_at	1.0
sku	1.0
price	1.0
qty_ordered	1.0
grand_total	1.0

Attribute	Completeness
increment_id	1.0
category_name_1	1.0
sales_commission_code	1.0
discount_amount	1.0
payment_method	1.0
Working Date	1.0
BI Status	1.0
MV	1.0
Year	1.0
Month	1.0
Customer Since	1.0
M-Y	1.0
FY	1.0
Customer ID	1.0
Unnamed: 21	0.0
Unnamed: 22	0.0
Unnamed: 23	0.0
Unnamed: 24	0.0
Unnamed: 25	0.0

KOBOLD.AI

Validity

Data validity describes the internal consistency of data types in an attribute. For example an attribute "Name" should only consist of Strings, while "Revenue" obviously should only populating numeric values.

In the following we examine the data type for each attribute. If there are outliers, i.e. very low percentages of string / numeric in a column when everything else is consistent, this might point to a data quality impairment and should be checked.

Attribute	Total	Empty	String	Numeric
item_id	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
status	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
created_at	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
sku	50000	5 (0.0%)	49995 (100.0%)	0 (0.0%)
price	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
qty_ordered	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
grand_total	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
increment_id	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
category_name_1	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
sales_commission_code	50000	0 (0.0%)	48798 (98.0%)	1202 (2.0%)
discount_amount	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)

KOBOLD.AI

Attribute	Total	Empty	String	Numeric
payment_method	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
Working Date	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
BI Status	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
MV	50000	0 (0.0%)	18402 (37.0%)	31598 (63.0%)
Year	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
Month	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
Customer Since	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
M-Y	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
FY	50000	0 (0.0%)	50000 (100.0%)	0 (0.0%)
Customer ID	50000	0 (0.0%)	0 (0.0%)	50000 (100.0%)
Unnamed: 21	50000	50000 (100.0%)	0 (0.0%)	0 (0.0%)
Unnamed: 22	50000	50000 (100.0%)	0 (0.0%)	0 (0.0%)
Unnamed: 23	50000	50000 (100.0%)	0 (0.0%)	0 (0.0%)
Unnamed: 24	50000	50000 (100.0%)	0 (0.0%)	0 (0.0%)
Unnamed: 25	50000	50000 (100.0%)	0 (0.0%)	0 (0.0%)

KOBOLD.AI

Duplicate rows

Duplicate rows are rows where all attributes are identical. In such cases, there is zero variability and therefore the duplicate rows need to be removed.

To extend the analysis, we also calculate duplicate rows with exclusion of one column at a time. This allows to identify cases in which a single column would be the only varying factor, but the rest of the columns are identical. Again, such cases need to be evaluated, as they do not provide any value.

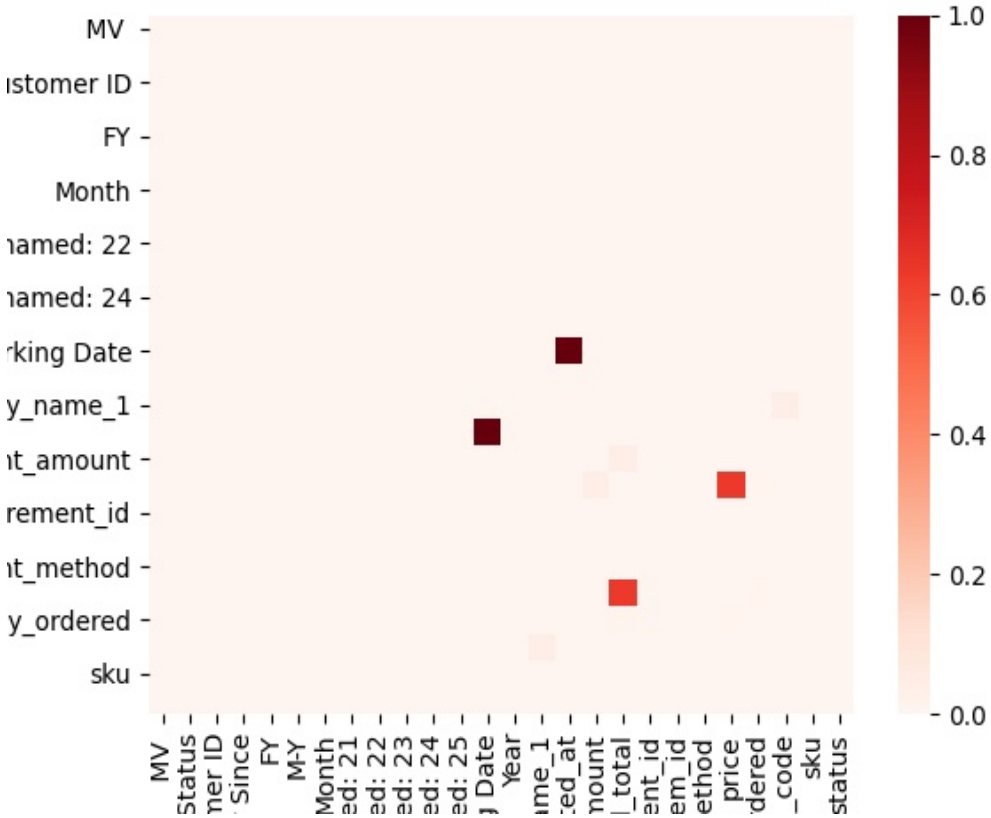
Attributes	Duplicate rows in percent
all	0.0
excluding item_id	0.0
excluding status	0.0
excluding created_at	0.0
excluding sku	0.0
excluding price	0.0
excluding qty_ordered	0.0
excluding grand_total	0.0
excluding increment_id	0.0
excluding category_name_1	0.0
excluding sales_commission_code	0.0
excluding discount_amount	0.0

KOBOLD.AI

Attributes	Duplicate rows in percent
excluding payment_method	0.0
excluding Working Date	0.0
excluding BI Status	0.0
excluding MV	0.0
excluding Year	0.0
excluding Month	0.0
excluding Customer Since	0.0
excluding M-Y	0.0
excluding FY	0.0
excluding Customer ID	0.0
excluding Unnamed: 21	0.0
excluding Unnamed: 22	0.0
excluding Unnamed: 23	0.0
excluding Unnamed: 24	0.0
excluding Unnamed: 25	0.0

KOBOLD.AI

Duplicate columns



A column is a duplicate of another column, if position and value of each cell are identical.

Duplicate columns are one of the most obvious indications to increase data sparsity by either merging two columns or removing one altogether, if it is identical to another column.

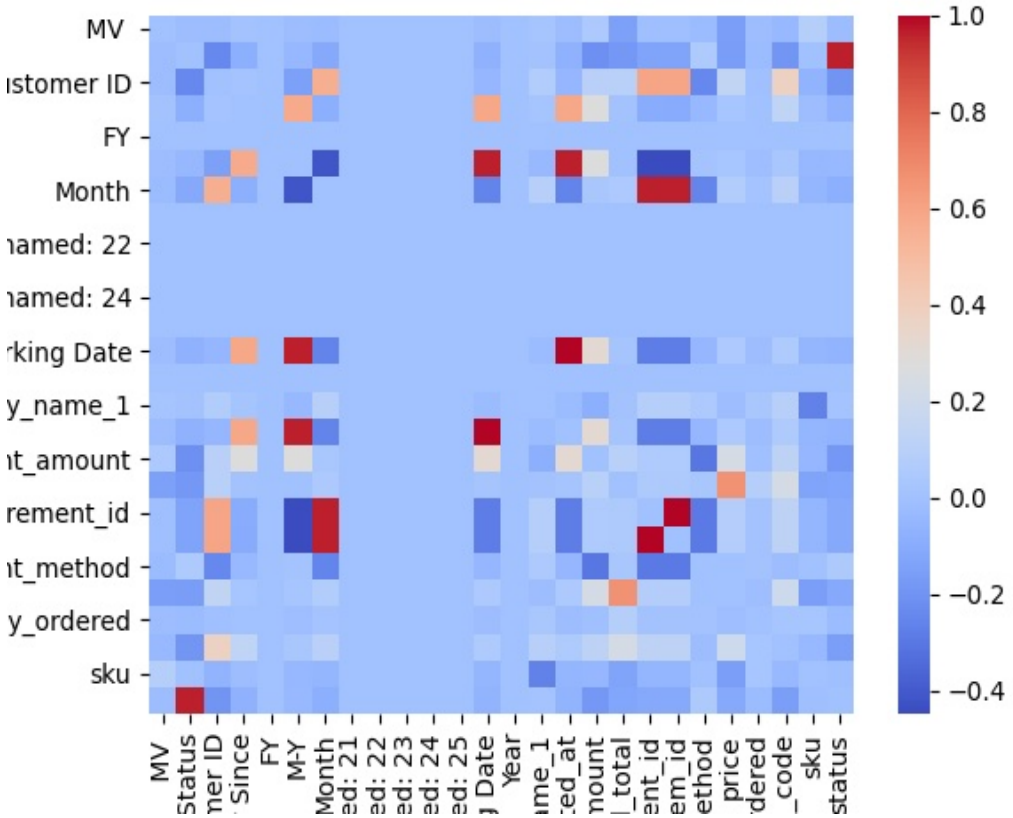
Attribute 1	Attribute 2	Duplicate content in percent
created_at	Working Date	1.0

KOBOLD.AI

Attribute 1	Attribute 2	Duplicate content in percent
Working Date	created_at	1.0

KOBOLD.AI

Data redundancy



Redundancy of two columns is calculated through the correlation of the content of those two attributes. If a high (positive or negative) correlation exists, especially if the correlation is 1 (or -1), the values of the columns are behaving exactly the same. Therefore, it has to be considered whether both attributes need to exist or one could be removed to achieve data sparseness.

The difference of data redundancy to duplicate columns is that the columns do not need to exhibit the exactly same values, but their content is behaving exactly the same. Therefore redundancy is more delicate and less detectable than identicial content.

Attribute 1	Attribute 2	Correlation (>0.8)
item_id	increment_id	1.0

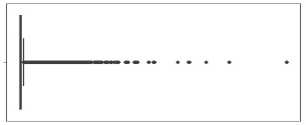
Attribute 1	Attribute 2	Correlation (>0.8)
item_id	Month	0.96
status	BI Status	0.96
created_at	Working Date	1.0
created_at	M-Y	0.96
increment_id	item_id	1.0
increment_id	Month	0.96
Working Date	created_at	1.0
Working Date	M-Y	0.96
BI Status	status	0.96
Month	item_id	0.96
Month	increment_id	0.96
M-Y	created_at	0.96
M-Y	Working Date	0.96

KOBOLD.AI

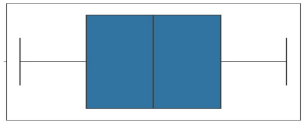
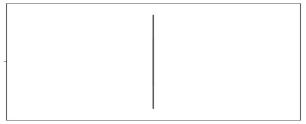
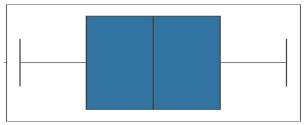
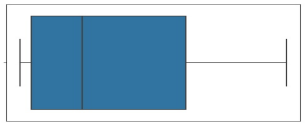
Data Accuracy (value distribution)

Data accuracy is concerned with the questions whether the values stored are correct. One initial way to identify wrong values is to check their distribution.

Below we have analysed the distribution of values of each numeric attribute. If the distribution consists of many outliers or has an uncommon distribution, inaccurate values might be present.

Column	Value distribution	Outlier percentage	Outlier examples
item_id		-	-
price		1.98%	51999.0, 44680.0, 70567.0, 62000.0...
qty_ordered		0.46%	6, 8, 8, 8...
grand_total		1.22%	83999.0, 72700.0, 96499.0, 87798.0...

KOBOLD.AI

Column	Value distribution	Outlier percentage	Outlier examples
increment_id		-	-
discount_amount		0.91%	6599.8, 5419.8, 3606.5, 10666.5...
Year		-	-
Month		-	-
Customer ID		-	-
Unnamed: 21		-	-

KOBOLD.AI

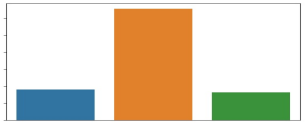
Column	Value distribution	Outlier percentage	Outlier examples
Unnamed: 22		-	-
Unnamed: 23		-	-
Unnamed: 24		-	-
Unnamed: 25		-	-

KOBOLD.AI

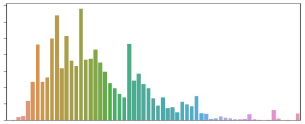
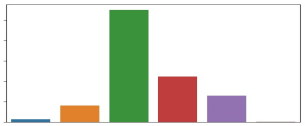
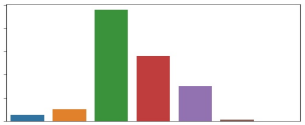
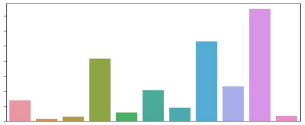
Data Accuracy (value length)

Data accuracy is concerned with the questions whether the values stored are correct. Another way to identify wrong values is to check their length.

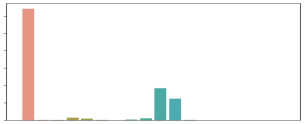
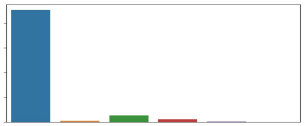
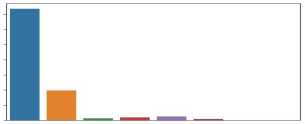
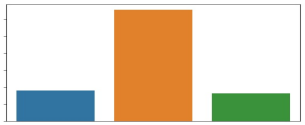
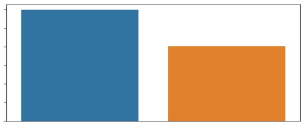
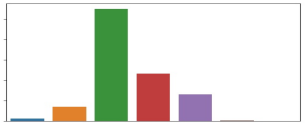
Below we have analysed the length of values of each attribute. If an attribute exhibits unusual lengths (e.g. an ISBN not being 10 digits long or having unusual low or high lengths), it is an indicator for bad data quality.

Column	Average length	Length range	Length Distribution
item_id	6.0	6 - 6	
status	8.7352	5 - 14	
created_at	8.98264	8 - 10	

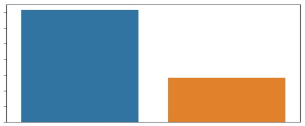
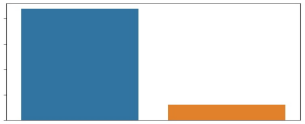
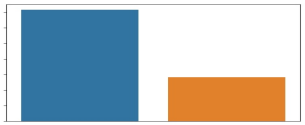
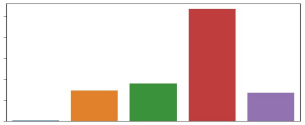
KOBOLD.AI

Column	Average length	Length range	Length Distribution
sku	22.59088	3 - 64	
price	5.37958	3 - 8	
qty_ordered	1.00204	1 - 3	
grand_total	5.50002	3 - 9	
increment_id	9.0	9 - 9	
category_name_1	12.29084	2 - 18	

KOBOLD.AI

Column	Average length	Length range	Length Distribution
sales_commission_code	5.16704	1 - 39	
discount_amount	3.21362	3 - 8	
payment_method	4.451	3 - 17	
Working Date	8.98264	8 - 10	
BI Status	3.80416	3 - 5	
MV	5.77606	3 - 11	

KOBOLD.AI

Column	Average length	Length range	Length Distribution
Year	4.0	4 - 4	
Month	1.2839	1 - 2	
Customer Since	6.122	6 - 7	
M-Y	6.2839	6 - 7	
FY	4.0	4 - 4	
Customer ID	3.65514	1 - 5	

Column	Average length	Length range	Length Distribution
Unnamed: 21	3.0	3 - 3	
Unnamed: 22	3.0	3 - 3	
Unnamed: 23	3.0	3 - 3	
Unnamed: 24	3.0	3 - 3	
Unnamed: 25	3.0	3 - 3