KPMG Data Quality Assessment

Albader H Qutub

Table of Contents

Introduction

This data quality assessment report aims to provide a comprehensive overview of our client's (Sprocket Central Pty Ltd) data quality landscape, addressing critical aspects such as Data Sources, Data Quality Framework, Data Quality Assessment Methodology, Data Quality Dimensions, and Findings & Recommendations. As we delve into each section, our objective is to assess the state of our data quality, identify potential areas of improvement, and offer actionable recommendations to enhance the overall reliability and utility of our data assets.

Data Sources

Our data quality assessment is centered around three primary datasets provided by the client:

- Customer Demographic Data

- Customer Addresses

- Transactions Data

Data Quality Framework

The Data Quality Framework is designed to uphold data quality across key dimensions:

Accuracy: We prioritize data precision through validation and cleansing.

Completeness: We ensure all relevant data is captured, addressing missing information.

Consistency: Standardized data entry and validation rules maintain uniformity.

Currency: Regular updates keep data current for real-time decisions.

Relevancy: Data sources are reviewed to maintain relevance to our operations.

Validity: Checks and validations enforce data conformity to defined criteria.

Uniqueness: Duplicate records are detected and resolved to preserve data integrity.

Tools and Software’s

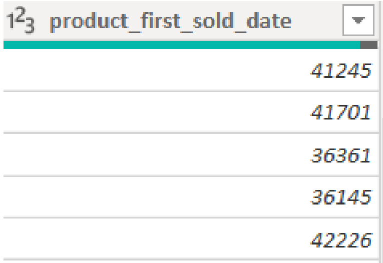
The data quality assessment is conducted using industry-standard tools and software, primarily Excel and Power BI. These widely recognized and accessible tools enable us to efficiently manage, analyze, and visualize large datasets, ensuring a thorough examination of data quality.

Findings & Recommendations

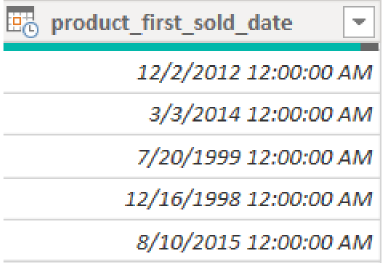
Transactions dataset:

- Dataset name confirmed

- After exploring the data type, it appears that (product_first_sold_date) column need to converted from integer to data-time

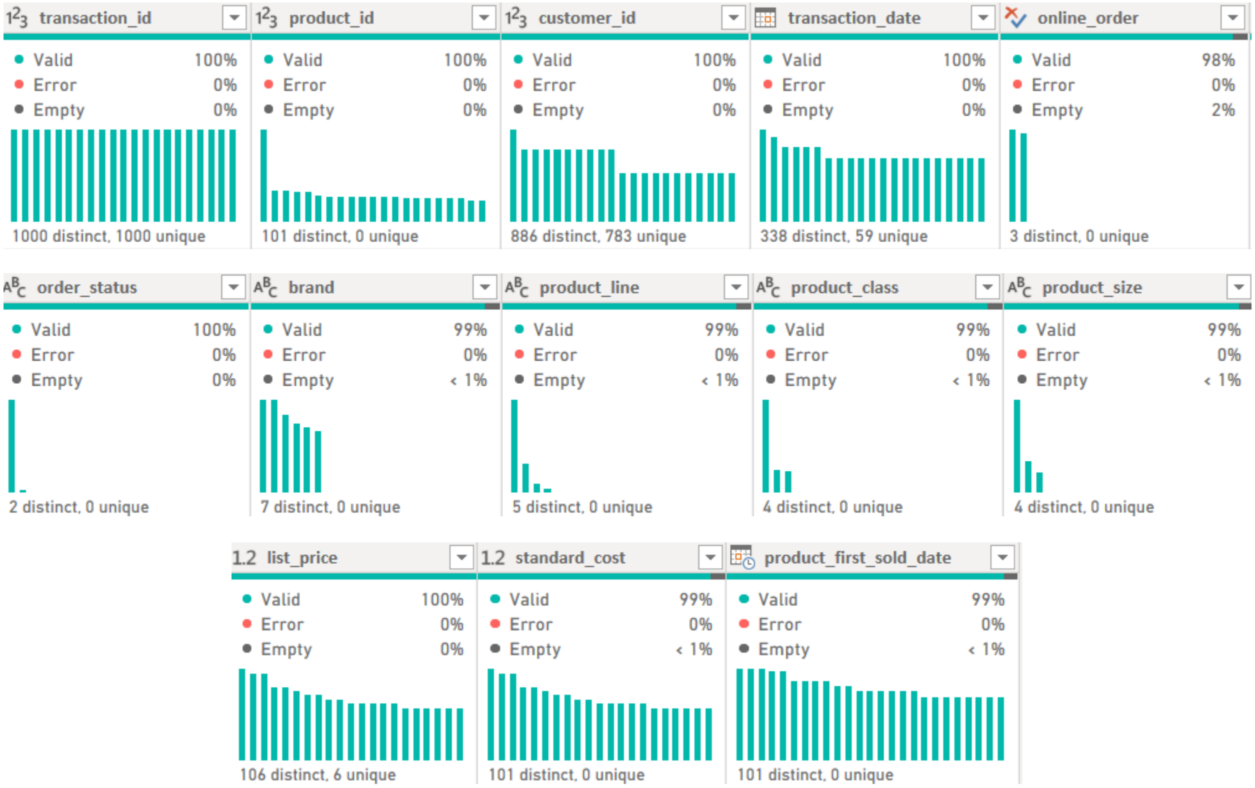


product_first_sold_date
41245
41701
36361
36145
42226



product_first_sold_date
12/2/2012 12:00:00 AM
3/3/2014 12:00:00 AM
7/20/1999 12:00:00 AM
12/16/1998 12:00:00 AM
8/10/2015 12:00:00 AM

- After exploring the columns, it appears that seven columns contain missing values, and their handling should align with the nature of the analysis, either through removal or appropriate treatment.



transaction_id
● Valid 100%
● Error 0%
● Empty 0%
1000 distinct, 1000 unique

product_id
● Valid 100%
● Error 0%
● Empty 0%
101 distinct, 0 unique

customer_id
● Valid 100%
● Error 0%
● Empty 0%
886 distinct, 783 unique

transaction_date
● Valid 100%
● Error 0%
● Empty 0%
338 distinct, 59 unique

online_order
● Valid 98%
● Error 0%
● Empty 2%
3 distinct, 0 unique

order_status
● Valid 100%
● Error 0%
● Empty 0%
2 distinct, 0 unique

brand
● Valid 99%
● Error 0%
● Empty < 1%
7 distinct, 0 unique

product_line
● Valid 99%
● Error 0%
● Empty < 1%
5 distinct, 0 unique

product_class
● Valid 99%
● Error 0%
● Empty < 1%
4 distinct, 0 unique

product_size
● Valid 99%
● Error 0%
● Empty < 1%
4 distinct, 0 unique

list_price
● Valid 100%
● Error 0%
● Empty 0%
106 distinct, 6 unique

standard_cost
● Valid 99%
● Error 0%
● Empty < 1%
101 distinct, 0 unique

product_first_sold_date
● Valid 99%
● Error 0%
● Empty < 1%
101 distinct, 0 unique

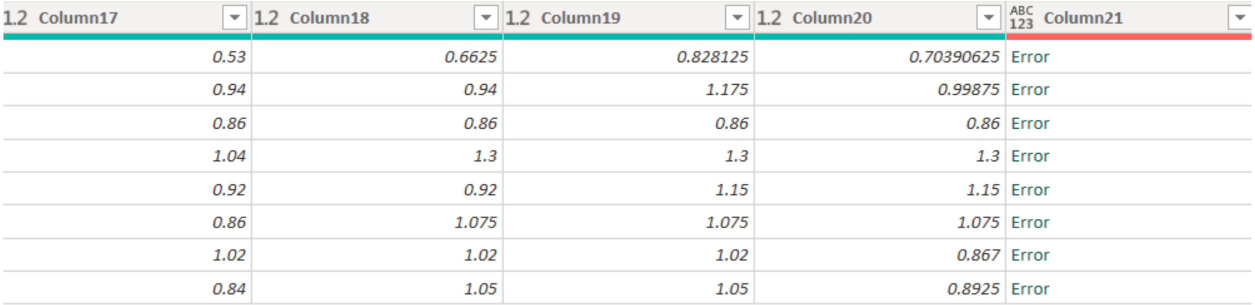
- There are no duplicate values.

New Customer List dataset:

- Dataset name confirmed

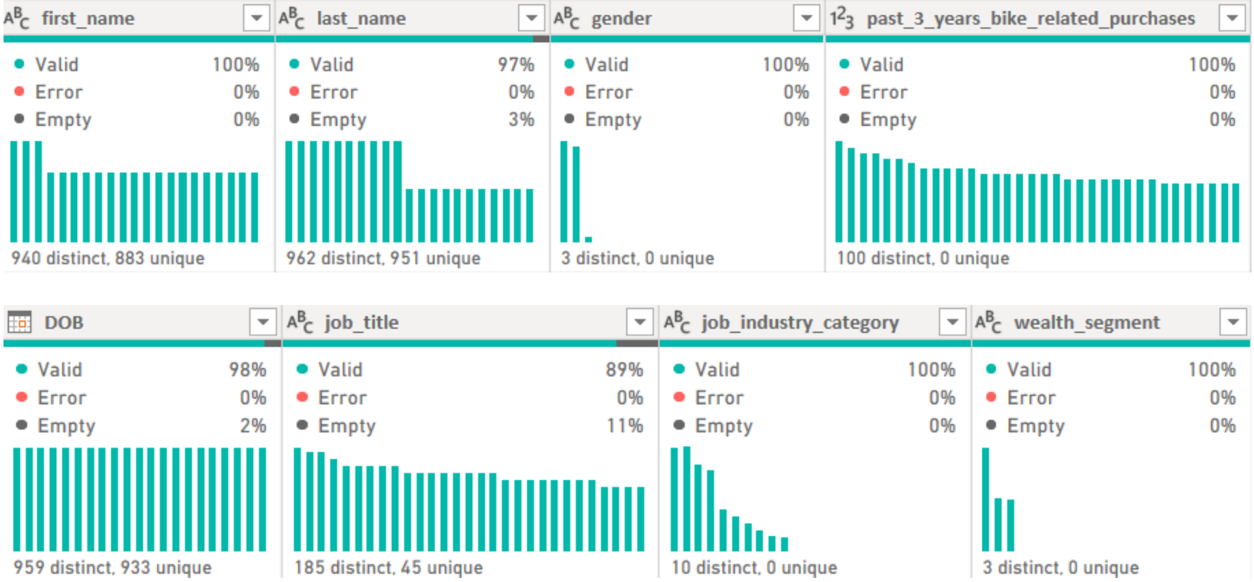
- Column data type confirmed

- After exploring the column, it appears that columns 17 to 21 are unnamed. These columns will be removed



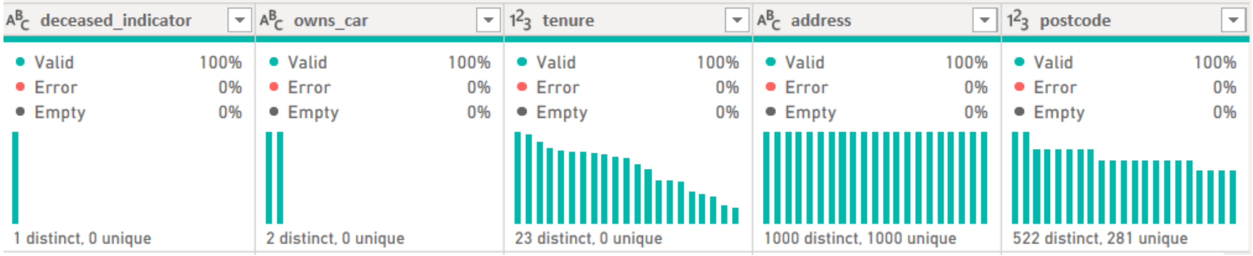
1.2 Column17	1.2 Column18	1.2 Column19	1.2 Column20	ABC 123 Column21
0.53	0.6625	0.828125	0.70390625	Error
0.94	0.94	1.175	0.99875	Error
0.86	0.86	0.86	0.86	Error
1.04	1.3	1.3	1.3	Error
0.92	0.92	1.15	1.15	Error
0.86	1.075	1.075	1.075	Error
1.02	1.02	1.02	0.867	Error
0.84	1.05	1.05	0.8925	Error

- After exploring the columns, it appears that four columns contain missing values, and their handling should align with the nature of the analysis, either through removal or appropriate treatment.

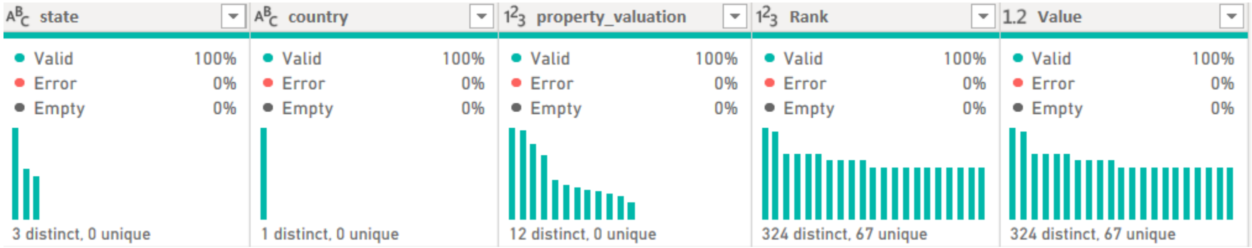


	first_name	last_name	gender	past_3_years_bike_related_purchases
● Valid	100%	97%	100%	100%
● Error	0%	0%	0%	0%
● Empty	0%	3%	0%	0%
	940 distinct, 883 unique	962 distinct, 951 unique	3 distinct, 0 unique	100 distinct, 0 unique

	DOB	job_title	job_industry_category	wealth_segment
● Valid	98%	89%	100%	100%
● Error	0%	0%	0%	0%
● Empty	2%	11%	0%	0%
	959 distinct, 933 unique	185 distinct, 45 unique	10 distinct, 0 unique	3 distinct, 0 unique



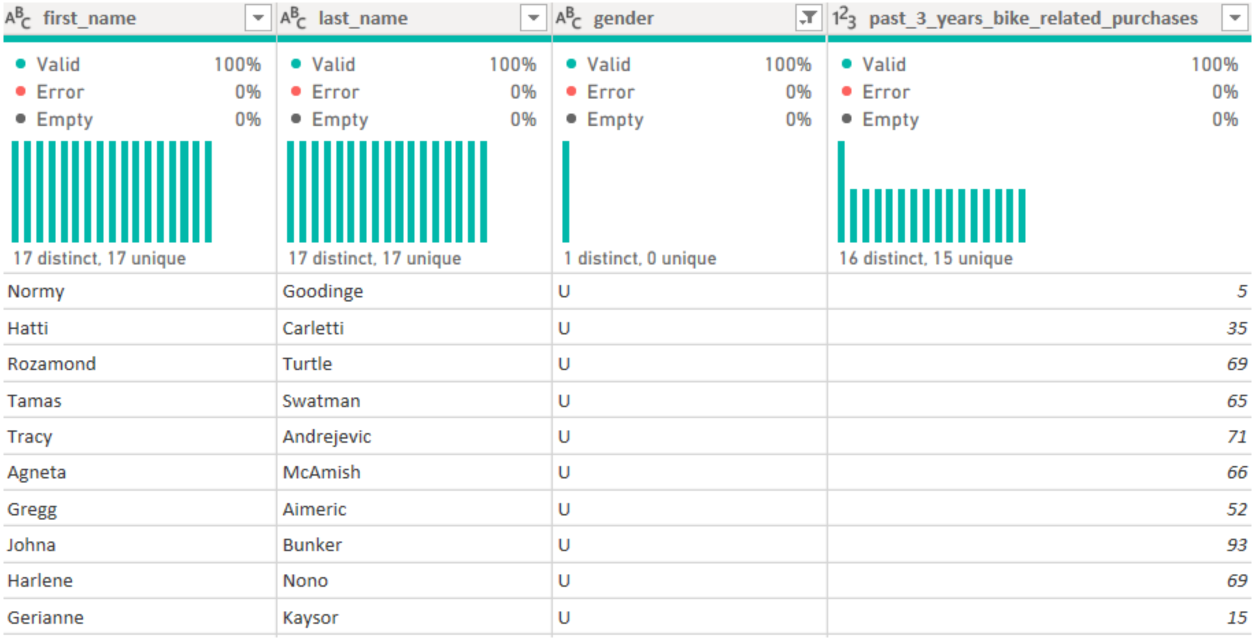
deceased_indicator	owns_car	tenure	address	postcode
● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%
● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%
● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%
1 distinct, 0 unique	2 distinct, 0 unique	23 distinct, 0 unique	1000 distinct, 1000 unique	522 distinct, 281 unique



state	country	property_valuation	Rank	Value
● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%
● Error 0%	● Error 0%	● Error 0%	● Error 0%	● Error 0%
● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%
3 distinct, 0 unique	1 distinct, 0 unique	12 distinct, 0 unique	324 distinct, 67 unique	324 distinct, 67 unique

- There are no duplicate values.

- After exploring the Gender column, it appears that there are 17 records with unknown/unspecified gender (U).



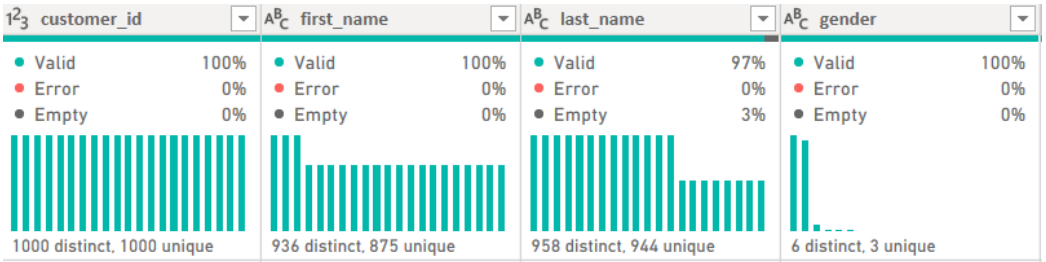
first_name	last_name	gender	past_3_years_bike_related_purchases
● Valid 100%	● Valid 100%	● Valid 100%	● Valid 100%
● Error 0%	● Error 0%	● Error 0%	● Error 0%
● Empty 0%	● Empty 0%	● Empty 0%	● Empty 0%
17 distinct, 17 unique	17 distinct, 17 unique	1 distinct, 0 unique	16 distinct, 15 unique
Normy	Goodinge	U	5
Hatti	Carletti	U	35
Rozamond	Turtle	U	69
Tamas	Swatman	U	65
Tracy	Andrejevic	U	71
Agneta	McAmish	U	66
Gregg	Aimeric	U	52
Johna	Bunker	U	93
Harlene	Nono	U	69
Gerianne	Kaysor	U	15

Customer Demographic Dataset:

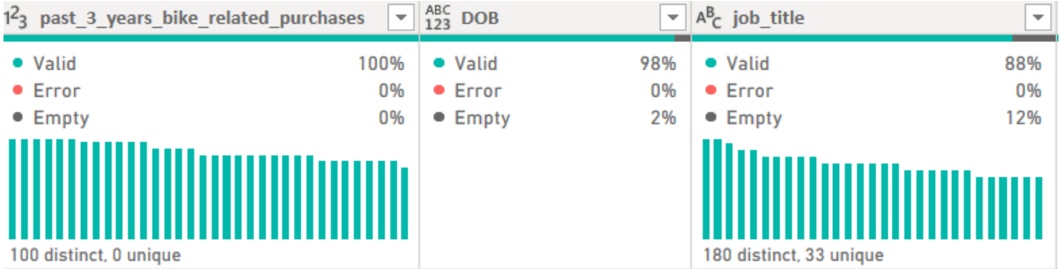
- Dataset name confirmed

- Column data type confirmed

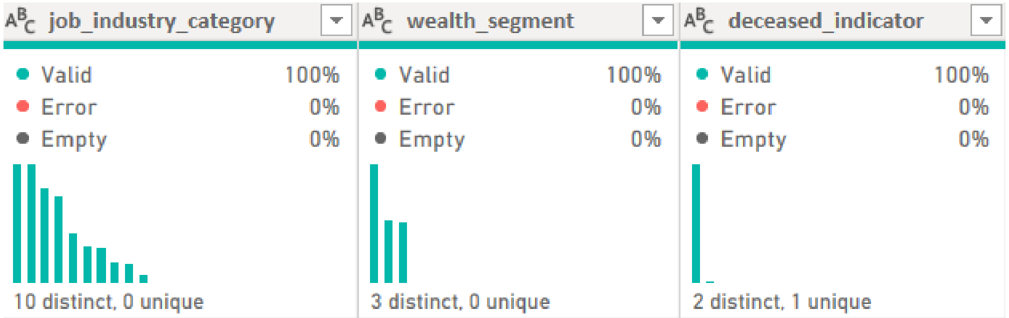
- After exploring the columns, it appears that five columns contain missing values, and their handling should align with the nature of the analysis, either through removal or appropriate treatment.



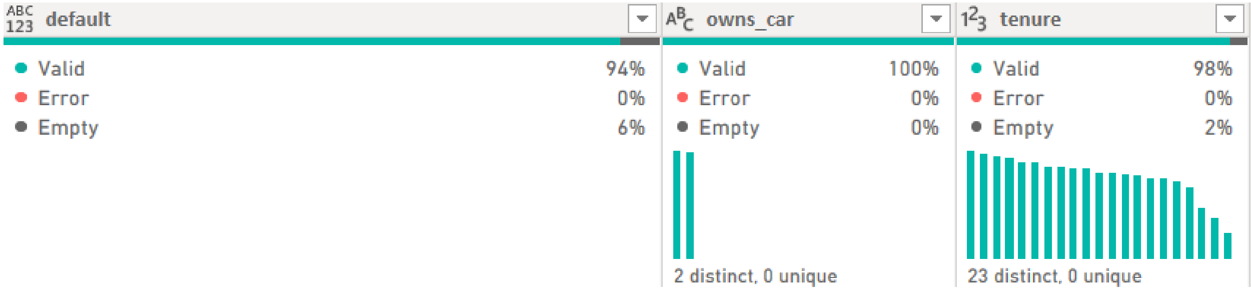
	customer_id	first_name	last_name	gender
● Valid	100%	100%	97%	100%
● Error	0%	0%	0%	0%
● Empty	0%	0%	3%	0%
	1000 distinct, 1000 unique	936 distinct, 875 unique	958 distinct, 944 unique	6 distinct, 3 unique



	past_3_years_bike_related_purchases	DOB	job_title
● Valid	100%	98%	88%
● Error	0%	0%	0%
● Empty	0%	2%	12%
	100 distinct, 0 unique		180 distinct, 33 unique

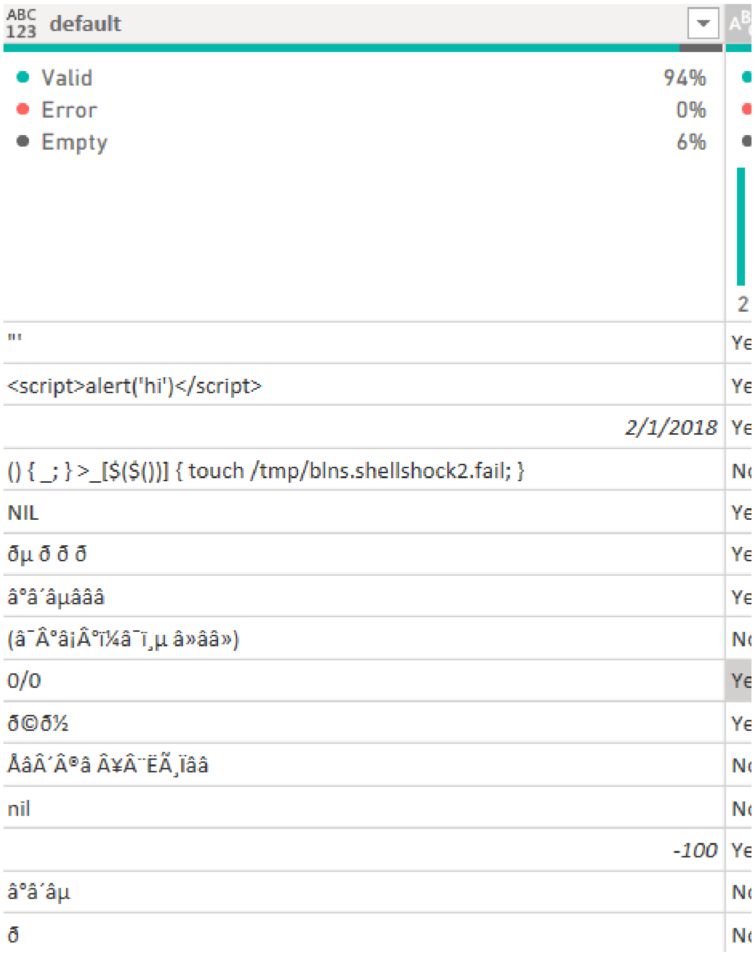


	job_industry_category	wealth_segment	deceased_indicator
● Valid	100%	100%	100%
● Error	0%	0%	0%
● Empty	0%	0%	0%
	10 distinct, 0 unique	3 distinct, 0 unique	2 distinct, 1 unique



	default	owns_car	tenure
● Valid	94%	100%	98%
● Error	0%	0%	0%
● Empty	6%	0%	2%
		2 distinct, 0 unique	23 distinct, 0 unique

- After exploring the (default) columns, it appears that the data is inconsistent and the column will be removed



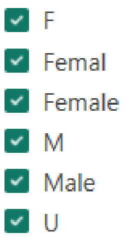
ABC 123 default

- ● Valid 94%
- ● Error 0%
- ● Empty 6%

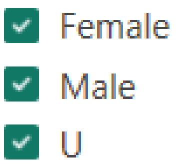
default	
"'	Ye
<script>alert('hi')</script>	Ye
2/1/2018	Ye
() { _; } >_[$($())] { touch /tmp/blns.shellshock2.fail; }	N
NIL	Ye
ðµ ð ð ð	Ye
â°â´âµâââ	Ye
(â¯Â°â¡Â°ï¼â¯ï¸µ â»ââ»)	N
0/0	Ye
ð©ð½	Ye
ÅâÂ´Âªâ Â¥Â¨ËÃ¸Ïââ	N
nil	N
-100	Ye
â°â´âµ	N
ð	N

- There are no duplicate values.

- The data in the (Gender) column is not consistent and will standardized



☑ F
☑ Femal
☑ Female
☑ M
☑ Male
☑ U



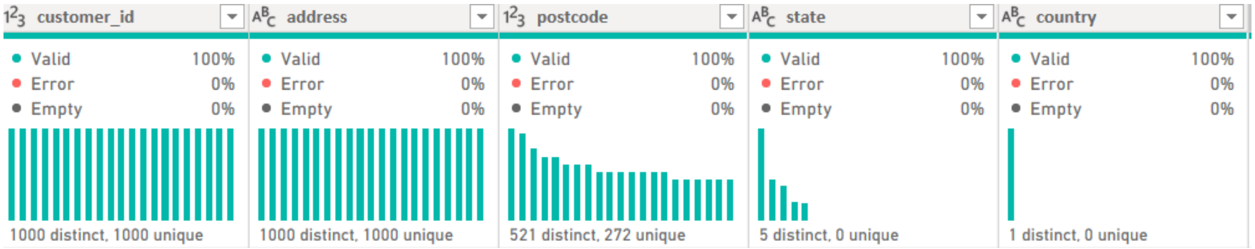
☑ Female
☑ Male
☑ U

Customer Address dataset:

- Dataset name confirmed

- Column data type confirmed

- No null values



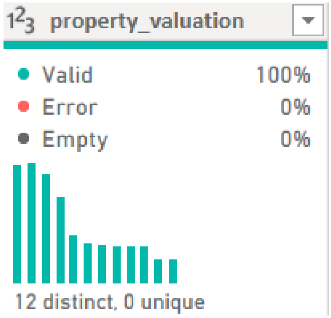
customer_id
Valid 100%
Error 0%
Empty 0%
1000 distinct, 1000 unique

address
Valid 100%
Error 0%
Empty 0%
1000 distinct, 1000 unique

postcode
Valid 100%
Error 0%
Empty 0%
521 distinct, 272 unique

state
Valid 100%
Error 0%
Empty 0%
5 distinct, 0 unique

country
Valid 100%
Error 0%
Empty 0%
1 distinct, 0 unique



property_valuation
Valid 100%
Error 0%
Empty 0%
12 distinct, 0 unique

- There are no duplicate values.

- All the columns appear to have consistent and correct information

Conclusion

In conclusion, the data quality assessment has provided a comprehensive view of our data's current state. It has revealed both strengths and areas for improvement across various dimensions, including accuracy, completeness, consistency, currency, relevancy, validity, and uniqueness. Addressing these findings will be crucial to enhancing data reliability and integrity.