# Data Quality Assessment (DQA) for improving IPEA's Growth Strategy

## Task

To build a Data Quality dashboard, where data attributes must be assessed along 6 dimensions. CDE's are defined to form rules, of which the status is shown in the DQ Dashboard.
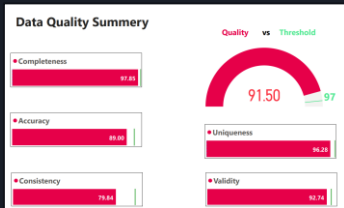
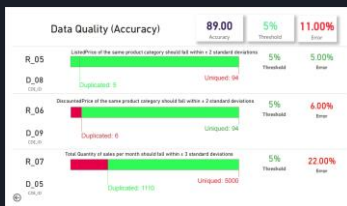## Dimensions of Data Quality

Completeness, Accuracy, Uniqueness, Consistency, Validity

## Approach

1-Defining Data Governance

2-Identifying Critical Data Elements

3-Designing Data Quality Rules

4-Setting thresholds

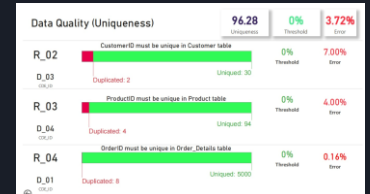5-Implementing DQ

6-Monitoring

## CDE & Rules

- Critical data elements consist of e.g., Customer ID, customer state.

- These CDE's are linked to Rules as e.g., if the Order ID is unique, if Customers are unique in the data table (but not in the order table).

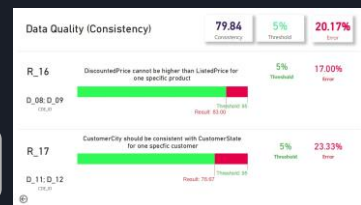| CDE Name | CDE Description | Table |
|---|---|---|
| Order ID | The unique identifier of a sales order | Order_Details (PK) |
| Order Date | The date when a sales order was placed | Order_Details |
| Customer ID | The unique identifier of a customer | Customer (PK); Order Details (FK) |
| Product ID | The unique identifier of a product | Product (PK); Order_Details (FK) |
| ... | ... | ... |

## Final result



Data Quality Summery

Completeness 97.85
Accuracy 89.00
Consistency 79.84
Uniqueness 96.28
Validity 92.74

Quality vs Threshold
91.50 / 97

## Completeness



Data Quality (Completeness) — 26K Total Rows — 97.85 Completeness — 0% Threshold — 2.15% Error

The CDE must be populated

R_01 — Completeness of CDEs

## Accuracy



Data Quality (Accuracy) — 89.00 Accuracy — 5% Threshold — 11.00% Error

R_05 D_08 — ListedPrice of the same product category should fall within ± 2 standard deviations — 5% Threshold — 5.00% Error — Duplicated: 5

R_06 D_09 — DiscountedPrice of the same product category fall within ± 2 standard deviations — 5% Threshold — 6.00% Error — Duplicated: 6

R_07 D_05 — Total Quantity of sales per month should fall within ± 2 standard deviations — 5% Threshold — 22.00% Error — Duplicated:1110 / Uniqued: 5000

## Uniqueness



Data Quality (Uniqueness) — 96.28 Uniqueness — 0% Threshold — 3.72% Error

R_02 D_03 — CustomerID must be unique in Customer table — 0% Threshold — 7.00% Error — Duplicated: 30 / Uniqued: 30

R_03 D_04 — ProductID must be unique in Product table — 0% Threshold — 4.00% Error — Duplicated: 4 / Uniqued: 94

R_04 D_01 — OrderID must be unique in Order_Details table — 0% Threshold — 0.16% Error — Duplicated: 8 / Uniqued: 5000

## Validity



Data Quality (Validity) — 92.78 — 4% Threshold — 7.23% Error

R_08 7.00% D_10 — Result: 93.00

R_09 0.80% D_02 — Customer Birth Date should not be less than 18 years or more than 120 years from now — Threshold: 95 — Result: 99.20

R_10 6.00% D_09 — Order Date cannot be a date in the future — Threshold: 94 — Result: 94.00

R_11 6.00% D_08 — Discounted Price cannot be a negative value — Threshold: 94 — Result: 94.00

R_12 0.00% D_10 — Listed Price cannot be a negative value — Result: 100.00

R_13 17.00% D_11 — Quantity cannot be a negative value — Threshold: 95 — Result: 83.00

R_14 20.00% D_12 — Customer City should be populated based on a pre specified list of strings — Threshold: 100 — Result: 80.00

R_15 1.00% D_07 — Customer State should be populated based on a pre specified list of strings — Threshold: 100 — Result: 99.00 — Product Category should be populated based on a pre specified list of strings

## consistency



Data Quality (Consistency) — 79.84 Consistency — 5% Threshold — 20.17% Error

R_16 D_08; D_09 — DiscountedPrice cannot be higher than ListedPrice for one specific product — 5% Threshold — 17.00% Error — Result: 83.00

R_17 D_11; D_12 — CustomerCity should be consistent with CustomerState for one specific customer — 5% Threshold — 23.33% Error — Result: 79.97

## Team:

Eissa Al-shari
Ali Ameen
Motea Al-omari
Saddam Obaid

## Supervisors:

Nick Bennink

Majda Kasmi

September 2023

Tools and software