Discoveries on the fifty database tables

# DB.table: AustralianFootball.players – Sports – Type: Real

pid (id):

Primary Key format: All 1911 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (10001 : 11911))

dob (datetime):

Datetime format: All 1911 datetime values are valid in the YYYYMMDD format in the range 1963/10/23 00:00 to 1996/12/11 00:00.

first\_name (name):

Name format: All 1911 first\_name values are valid.

Frequency Distribution:

first\_name Frequency

Matthew 52

Andrew 43

Ben 43

Tom 42

Daniel 41

Luke 36

Michael 36

Josh 33

Sam 33

Adam 32

... ...

Tory 1

Touk 1

Tristen 1

Viv 1

Wade 1

Warren 1

Winston 1

Zaine 1

Zak 1

Zephaniah 1

Range of Values: (Aaron to Zephaniah)

last\_name (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

29 Capitalization/Format issue(s) at index(es): [(282, 'LeCras'), (310, 'OFarrell'), (311, 'OHanlon'), (318, 'OBrien'), (355, 'OShea'), (440, 'DellOlio'), (456, 'OBrien'), (524, 'OBrien'), (759, 'OMeara'), (809, 'OReilly'), ('...', '...'), (1504, 'OLoughlin'), (1512, 'AhMat'), (1562, 'OConnor'), (1563, 'OKeefe'), (1630, 'OKeeffe'), (1635, 'OhAilpin'), (1647, 'OBree'), (1702, 'OReilly'), (1765, 'OBrien'), (1902, 'OBrien')] (displaying only the first and last 10 items)

Frequency Distribution:

last\_name Frequency

Smith 20

Johnson 15

Brown 13

Jones 12

White 12

Clarke 9

Williams 9

Campbell 8

Hall 8

OBrien 8

... ...

Yarran 1

Yeo 1

Yze 1

Zaharakis 1

Zantuck 1

Ziebell 1

Zorko 1

de Boer 1

de Goey 1

van Unen 1

Range of Values: (Ablett to van Unen)

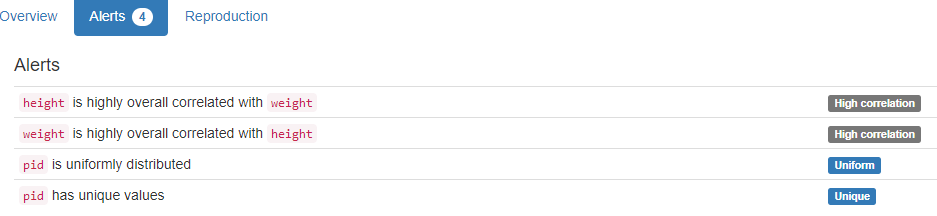
height:

Numerical >=0 format: All 1911 values are numerical and greater or equal to 0 in the range (163:211).

weight:

Numerical >=0 format: All 1911 values are numerical and greater or equal to 0 in the range (63:120).

YDATA Profiling results: It just found out that the PK column 'pid' has unique values and is uniformly distributed:



# DB.table: AustralianFootball.matches – Sports – Type: Real

mid (id):

Primary Key format: All 3050 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (2000\_EF\_102\_118 : 2015\_R23\_115\_117)

season:

All 3050 values are correctly categorical.

Categorical format with 16 unique values:

Category Frequency

2012 207

2013 207

2014 207

2015 197

2011 196

2010 186

2000 185

2001 185

2002 185

2003 185

2004 185

2005 185

2006 185

2007 185

2008 185

2009 185

round (varchar):

String format: All 3050 string values are valid.

String range (lexicographical): (EF : SF)

tid1 (id):

Foreign Key format: All 3050 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (101 : 117)

References: teams.tid

tid2 (id):

Foreign Key format: All 3050 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (102 : 118)

References: teams.tid

tid1\_loc (location):

All 3050 values are correctly categorical.

Categorical format with 2 unique values:

Category Frequency

h 1527

a 1523

tid2\_loc (location):

All 3050 values are correctly categorical.

Categorical format with 2 unique values:

Category Frequency

a 1527

h 1523

start\_dt (datetime):

Datetime format: All 3050 datetime values are valid in the YYYYMMDD format in the range 2000/03/08 19:40 to 2015/09/06 15:20.

venue:

All 3050 values are correctly categorical.

Categorical format with 21 unique values:

Category Frequency

M.C.G. 784

Docklands 751

Subiaco 358

Football Park 325

Gabba 194

S.C.G. 147

Kardinia Park 122

Carrara 63

York Park 56

Stadium Australia 53

... ...

Princes Park 44

Manuka Oval 37

Sydney Showground 29

Marrara Oval 16

Bellerive Oval 9

W.A.C.A. 6

Cazaly's Stadium 5

Wellington 3

Traeger Park 2

Blacktown 1

att (integer):

Numerical format: All 3050 values are numerical in the range (4866:100016).

margin (integer):

Numerical format: All 3050 values are numerical in the range (0:186).

win\_score (score):

Numerical >=0 format: All 3050 values are numerical and greater or equal to 0 in the range (48:233).

win\_tid (integer):

Numerical format: All 3050 values are numerical in the range (100:118).

win\_loc (location):

All 3050 values are correctly categorical.

Categorical format with 3 unique values:

Category Frequency

h 1756

a 1266

d 28

lose\_score (score):

Numerical >=0 format: All 3050 values are numerical and greater or equal to 0 in the range (13:159).

lose\_loc (location):

All 3050 values are correctly categorical.

Categorical format with 3 unique values:

Category Frequency

a 1756

h 1266

d 28

lose\_tid (integer):

Numerical format: All 3050 values are numerical in the range (100:118).

h\_tid (integer):

Numerical format: All 3050 values are numerical in the range (101:118).

h\_score (score):

Numerical format: All 3050 values are numerical in the range (21:233).

h\_q1g (integer):

Numerical format: All 3050 values are numerical in the range (0:15).

h\_q2g (integer):

Numerical format: All 3050 values are numerical in the range (0:20).

h\_q3g (integer):

Numerical format: All 3050 values are numerical in the range (1:28).

h\_q4g (integer):

Numerical format: All 3050 values are numerical in the range (3:37).

h\_etg (integer):

Numerical format: All 3050 values are numerical in the range (0:10).

h\_q1b (integer):

Numerical format: All 3050 values are numerical in the range (0:17).

h\_q2b (integer):

Numerical format: All 3050 values are numerical in the range (1:24).

h\_q3b (integer):

Numerical format: All 3050 values are numerical in the range (2:27).

h\_q4b (integer):

Numerical format: All 3050 values are numerical in the range (4:118).

h\_etb (integer):

Numerical format: All 3050 values are numerical in the range (0:222).

a\_tid (integer):

Numerical format: All 3050 values are numerical in the range (0:118).

a\_score (score):

Numerical >=0 format: All 3050 values are numerical and greater or equal to 0 in the range (0:164).

a\_q1g (integer):

Numerical format: All 3050 values are numerical in the range (0:29).

a\_q2g (integer):

Numerical format: All 3050 values are numerical in the range (1:35).

a\_q3g (integer):

Numerical format: All 3050 values are numerical in the range (0:20).

a\_q4g (integer):

Numerical format: All 3050 values are numerical in the range (0:26).

a\_etg (integer):

Numerical format: All 3050 values are numerical in the range (0:21).

a\_q1b (integer):

Numerical format: All 3050 values are numerical in the range (0:25).

a\_q2b (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2914 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (15, 'NULL'), (16, 'NULL'), ('...', '...'), (3040, 'NULL'), (3041, 'NULL'), (3042, 'NULL'), (3043, 'NULL'), (3044, 'NULL'), (3045, 'NULL'), (3046, 'NULL'), (3047, 'NULL'), (3048, 'NULL'), (3049, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:11.0).

a\_q3b (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):   
 2914 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (15, 'NULL'), (16, 'NULL'), ('...', '...'), (3040, 'NULL'), (3041, 'NULL'), (3042, 'NULL'), (3043, 'NULL'), (3044, 'NULL'), (3045, 'NULL'), (3046, 'NULL'), (3047, 'NULL'), (3048, 'NULL'), (3049, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:16.0).

a\_q4b (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):  
 2914 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (15, 'NULL'), (16, 'NULL'), ('...', '...'), (3040, 'NULL'), (3041, 'NULL'), (3042, 'NULL'), (3043, 'NULL'), (3044, 'NULL'), (3045, 'NULL'), (3046, 'NULL'), (3047, 'NULL'), (3048, 'NULL'), (3049, 'NULL')] (displaying only the first and last 10 items)

Range of values: (2.0:19.0).

a\_etb (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2914 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (15, 'NULL'), (16, 'NULL'), ('...', '...'), (3040, 'NULL'), (3041, 'NULL'), (3042, 'NULL'), (3043, 'NULL'), (3044, 'NULL'), (3045, 'NULL'), (3046, 'NULL'), (3047, 'NULL'), (3048, 'NULL'), (3049, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:15.0).

target:

All 3050 values are correctly categorical.

Categorical format with 2 unique values:

Category Frequency

0 1529

1. 1521

YDATA Profiling results: It found the same 4 Missing values columns because they contain NULLs. It also found that the PK column 'mid' is unique.

A white rectangular object with a black border

Description automatically generated

…



# DB.table: Biodegradability.molecule – Medicine – Type: Real

Start of Analysis on: 2024-09-03 12:05:24

molecule\_id (id):

Primary Key format: All 328 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (i100\_02\_7i : i99\_65\_0i)

activity (float):

Numerical format: All 328 values are numerical in the range (2.14007:11.2742).

logp (float):

Numerical format: All 328 values are numerical in the range (-3.81:12.11).

mweight (weight):

Numerical >=0 format: All 328 values are numerical and greater or equal to 0 in the range (27.0259:1019.64).

YDATA Profiling results: It just found out that the PK column 'molecule\_id' has unique values:

A white background with black lines

Description automatically generated

# DB.table: ccs.transactions\_1k – Finance – Type: Real

Start of Analysis on: 2024-09-03 14:45:14

TransactionID (id):

Primary Key format: All 1000 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 1000)

Date:

Date format: All 1000 date values are valid in the YYYYMMDD format in the range 2012-08-23 to 2012-08-26.

Time:

Time format: All 1000 time values are valid in the range 00:07:00 to 23:20:00.

CustomerID (id):

Foreign Key format: All 1000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (96 : 49838)

References: customers.CustomerID

CardID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

182 Duplicate value(s) at index(es): [[1, 550134], [2, 550134], [7, 729387], [8, 729387], [25, 753582], [26, 753582], [47, 664247], [48, 664247], [54, 688291], [55, 688291], ('...', '...'), [951, 483052], [952, 483052], [959, 626787], [960, 626787], [965, 442969], [966, 199529], [975, 523936], [976, 523936], [977, 501247], [978, 501247]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

98 Uniqueness violation(s) at index(es): [(2, 550134), (8, 729387), (26, 753582), (48, 664247), (55, 688291), (78, 450132), (84, 568944), (85, 568944), (86, 568944), (95, 660091), ('...', '...'), (877, 562088), (906, 199431), (915, 418233), (943, 65851), (952, 483052), (960, 626787), (965, 442969), (966, 199529), (976, 523936), (978, 501247)] (displaying only the first and last 10 items)

Numeric range of values: (26228 : 775970)

GasStationID (id):

Foreign Key format: All 1000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (48 : 5481)

References: gasstations.GasStationID

ProductID (id):

Foreign Key format: All 1000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (2 : 352)

References: products.ProductID

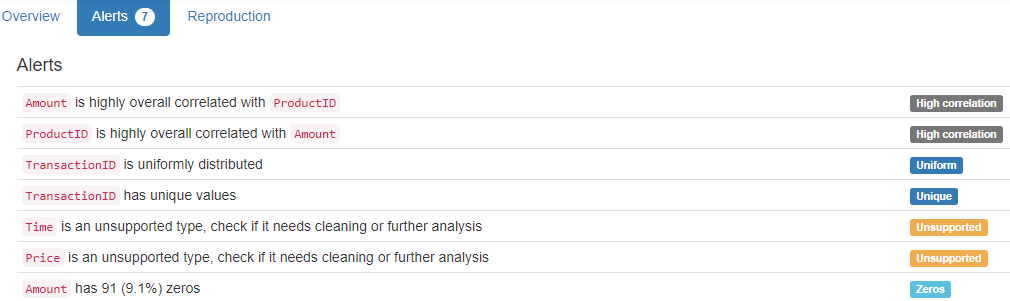
Amount:

Numerical format: All 1000 values are numerical in the range (0:264).

Price:

Numerical format: All 1000 values are numerical in the range (1.76:5762.49).

YDATA Profiling results: It just found out that the PK column 'TransactionID' has unique values and is uniformly distributed. It also shows that cannot support columns Time and Price.



# DB.table: classicmodels.customers – Retail – Type: Synth

customerNumber (id):

Primary Key format: All 122 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (103 : 496)

customerName (name):

Name format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

2 Extraneous data value(s) at index(es): [(35, 'Toys4GrownUps.com'), (84, 'Gifts4AllAges.com')]

DQI #15 (Domain Violation - Accuracy):

6 Capitalization/Format issue(s) at index(es): [(0, 'Atelier graphique'), (23, "La Corne D'abondance, Co."), (46, 'giftsbymail.co.uk'), (82, 'GiftsForHim.com'), (92, "L'ordine Souveniers"), (108, 'FunGiftIdeas.com')]

Frequency Distribution:

customerName Frequency

ANG Resellers 1

AV Stores, Co. 1

Alpha Cognac 1

American Souvenirs Inc 1

Amica Models & Co. 1

Anna's Decorations, Ltd 1

Anton Designs, Ltd. 1

Asian Shopping Network, Co 1

Asian Treasures, Inc. 1

Atelier graphique 1

... ...

Toms Spezialitäten, Ltd 1

Toys of Finland, Co. 1

Toys4GrownUps.com 1

UK Collectables, Ltd. 1

Vida Sport, Ltd 1

Vitachrome Inc. 1

Volvo Model Replicas, Co 1

Warburg Exchange 1

West Coast Collectables Co. 1

giftsbymail.co.uk 1

Range of Values: (ANG Resellers to giftsbymail.co.uk)

contactLastName (name):

Name format: All 122 contactLastName values are valid.

Frequency Distribution:

contactLastName Frequency

Young 4

Brown 3

Frick 3

Franco 2

Hernandez 2

King 2

Murphy 2

Nelson 2

Taylor 2

Thompson 2

... ...

Sommer 1

Suominen 1

Tamuri 1

Tonini 1

Tseng 1

Urs 1

Victorino 1

Walker 1

Yoshido 1

de Castro 1

Range of Values: (Accorti to de Castro)

contactFirstName (name):

Name format: All 122 contactFirstName values are valid.

Frequency Distribution:

contactFirstName Frequency

Julie 4

Alexander 2

Juri 2

Leslie 2

Michael 2

Mike 2

Steve 2

Sue 2

Valarie 2

Adrian 1

... ...

Tony 1

Veysel 1

Violeta 1

Wales 1

Wendy 1

William 1

Wing 1

Yoshi 1

Yu 1

Zbyszek 1

Range of Values: (Adrian to Zbyszek )

phone:

Phone format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

11 Incorrect telephone number format issue(s) at index(es): [(0, '40.32.2555'), (3, '40.67.8555'), (13, '78.32.5555'), (22, '20.16.1555'), (23, '(1) 42.34.2555'), (37, '88.60.1555'), (47, '61.77.6555'), (51, '30.59.8555'), (79, '91.24.4555'), (80, '26.47.1555'), (94, '(1) 47.55.6555')]

addressLine1 (address):

Street format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

11 Capitalization/Format issue(s) at index(es): [(4, 'Erling Skakkes gate 78'), (6, 'ul. Filtrowa 68'), (21, 'Estrada da saúde n. 58'), (22, '184, chaussée de Tournai'), (23, '265, boulevard Charonne'), (37, '24, place Kléber'), (51, "67, avenue de l'Europe"), (56, 'Via Ludovico il Moro 22'), (80, "59 rue de l'Abbaye"), (86, 'Jardim das rosas n. 32'), (109, 'c/ Gobelas, 19-1 Urb. La Florida')]

addressLine2 (address):

Street format: Error(s) found:

DQI #1 (Missing Data - Completeness):

100 Blank/Empty/Null/Missing value(s) at index(es): [(0, NULL), (1, NULL), (3, NULL), (4, NULL), (5, NULL), (6, NULL), (7, NULL), (8, NULL), (9, NULL), (10, NULL), ('...', '...'), (112, NULL), (113, NULL), (114, NULL), (115, NULL), (116, NULL), (117, NULL), (118, NULL), (119, NULL), (120, NULL), (121, NULL)] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

1 Capitalization/Format issue(s) at index(es): [(91, 'NatWest Center #13-03')]

city:

City format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

2 Capitalization/Format issue(s) at index(es): [(26, 'Kita-ku'), (93, 'Minato-ku')]

Frequency Distribution (showing top and bottom 10 of 96 cities):

City Frequency

Madrid 5

NYC 5

Brickhaven 3

Paris 3

Singapore 3

Auckland 2

Boston 2

Cambridge 2

Frankfurt 2

Glendale 2

... ...

Stuttgart 1

Torino 1

Toulouse 1

Tsawassen 1

Vancouver 1

Versailles 1

Warszawa 1

Wellington 1

White Plains 1

Århus 1

state:

State format: Error(s) found:

DQI #1 (Missing Data - Completeness):

73 Blank/Empty/Null/Missing value(s) at index(es): [(0, NULL), (3, NULL), (4, NULL), (6, NULL), (7, NULL), (10, NULL), (11, NULL), (12, NULL), (13, NULL), (14, NULL), ('...', '...'), (106, NULL), (107, NULL), (109, NULL), (111, NULL), (113, NULL), (114, NULL), (115, NULL), (116, NULL), (119, NULL), (121, NULL)] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 19 states/provinces):

State Frequency

NULL 73

CA 11

MA 9

NY 6

CT 4

PA 3

BC 2

NSW 2

Victoria 2

Co. Cork 1

Isle of Wight 1

NH 1

NJ 1

NV 1

Osaka 1

Pretoria 1

Queensland 1

Québec 1

Tokyo 1

postalCode:

Postal Code format: Error(s) found:

DQI #1 (Missing Data - Completeness):

7 Blank/Empty/Null/Missing value(s) at index(es): [(38, NULL), (69, NULL), (78, NULL), (82, NULL), (96, NULL), (111, NULL), (121, NULL)]

DQI #15 (Domain Violation - Accuracy):

1 Short length alphanumeric value(s) at index(es): [(30, '2')]

country:

Country format: All 122 country values are valid.

Frequency Distribution (showing top and bottom 10 of 27 countries):

Country Frequency

USA 36

Germany 13

France 12

Spain 7

Australia 5

UK 5

Italy 4

New Zealand 4

Canada 3

Finland 3

... ...

Japan 2

Portugal 2

Sweden 2

Hong Kong 1

Israel 1

Netherlands 1

Philippines 1

Poland 1

Russia 1

South Africa 1

salesRepEmployeeNumber (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

22 Blank/Empty/Null/Missing value(s) at index(es): [(6, NULL), (21, NULL), (36, NULL), (41, NULL), (44, NULL), (48, NULL), (54, NULL), (59, NULL), (62, NULL), (63, NULL), ('...', '...'), (81, NULL), (83, NULL), (86, NULL), (95, NULL), (99, NULL), (107, NULL), (109, NULL), (113, NULL), (114, NULL), (115, NULL)] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

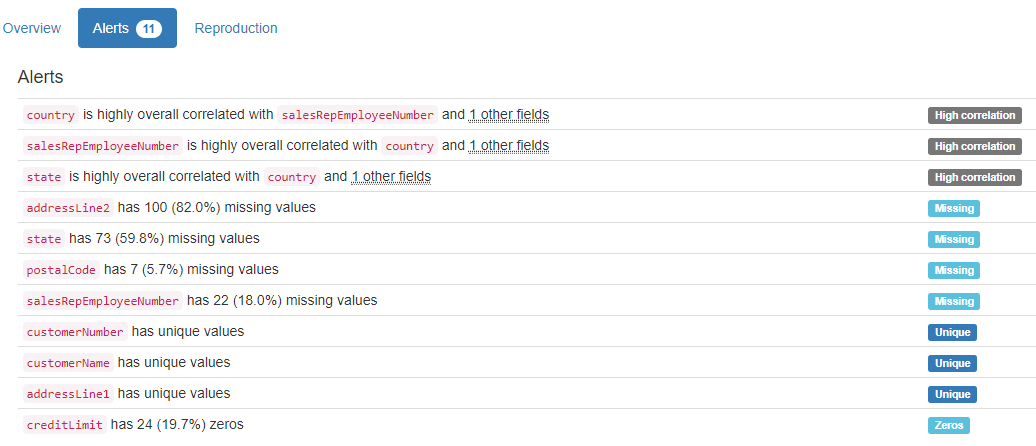
Numeric range of values: (1165.0 : 1702.0)

References: employees.employeeNumber

creditLimit (credit):

Numerical format: All 122 values are numerical in the range (0.0:227600.0).

YDATA Profiling results: It found the same 4 Missing values columns because they contain NULLs. It also found that the PK column 'customerNumber' is Unique.



# DB.table: Countries.Metadata – Countries – Geography – Type: Real

Country Code (id):

Primary Key format: All 247 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (ABW : ZWE)

Country Name (country):

Country format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

25 Capitalization/Format issue(s) at index(es): [(39, "Côte d'Ivoire"), (47, 'Caribbean small states'), (60, 'East Asia & Pacific (all income levels)'), (62, 'Europe & Central Asia (all income levels)'), (65, 'Euro area'), (71, 'Fragile and conflict affected situations'), (91, 'High income'), (94, 'Heavily indebted poor countries (HIPC)'), (125, 'Latin America & Caribbean (all income levels)'), (126, 'Least developed countries: UN classification'), ('...', '...')]

Frequency Distribution (showing top and bottom 10 of 247 countries):

Country Frequency

Afghanistan 1

Albania 1

Algeria 1

American Samoa 1

Andorra 1

Angola 1

Antigua and Barbuda 1

Arab World 1

Argentina 1

Armenia 1

... ...

Uzbekistan 1

Vanuatu 1

Venezuela, RB 1

Vietnam 1

Virgin Islands (U.S.) 1

West Bank and Gaza 1

World 1

Yemen, Rep. 1

Zambia 1

Zimbabwe 1

Region:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

33 Blank/Empty/Null/Missing value(s) at index(es): [(5, NULL), (34, NULL), (47, NULL), (59, NULL), (60, NULL), (61, NULL), (62, NULL), (65, NULL), (70, NULL), (71, NULL), ('...', '...'), (171, NULL), (172, NULL), (174, NULL), (187, NULL), (193, NULL), (204, NULL), (206, NULL), (207, NULL), (232, NULL), (241, NULL)] (displaying only the first and last 10 items)

Categorical format with 8 unique values:

Category Frequency

Europe & Central Asia 57

Sub-Saharan Africa 48

Latin America & Caribbean 41

East Asia & Pacific 36

NULL 33

Middle East & North Africa 21

South Asia 8

North America 3

IncomeGroup (group):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

33 Blank/Empty/Null/Missing value(s) at index(es): [(5, NULL), (34, NULL), (47, NULL), (59, NULL), (60, NULL), (61, NULL), (62, NULL), (65, NULL), (70, NULL), (71, NULL), ('...', '...'), (171, NULL), (172, NULL), (174, NULL), (187, NULL), (193, NULL), (204, NULL), (206, NULL), (207, NULL), (232, NULL), (241, NULL)] (displaying only the first and last 10 items)

Categorical format with 6 unique values:

Category Frequency

Upper middle income 55

Lower middle income 50

High income: nonOECD 44

Low income 34

NULL 33

High income: OECD 31

SpecialNotes (note):

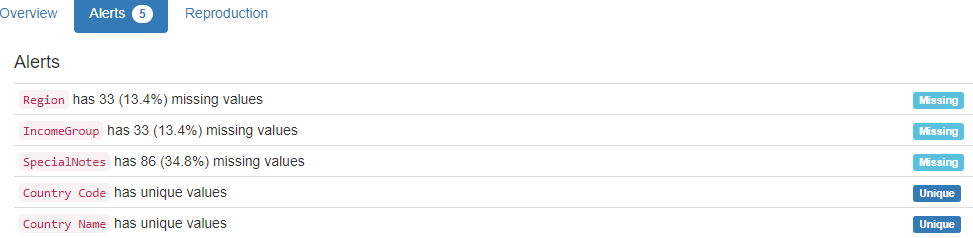
String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

86 Blank/Empty/Null/Missing value(s) at index(es): [(3, NULL), (4, NULL), (7, NULL), (8, NULL), (9, NULL), (14, NULL), (16, NULL), (17, NULL), (21, NULL), (23, NULL), ('...', '...'), (214, NULL), (217, NULL), (221, NULL), (225, NULL), (227, NULL), (229, NULL), (231, NULL), (235, NULL), (237, NULL), (238, NULL)] (displaying only the first and last 10 items)

String range (lexicographical): (2013 national accounts estimates are based on IMF reports. Based on official government statistics, value added is measured at producer prices through 1997 and at basic prices from 1998 onward. : World aggregate.)

YDATA Profiling results: It found the same 3 Missing values columns because they contain NULLs. It also found that the PK column 'Country Code' is Unique.



# DB.table: Credit.charge – Retail – Type: Synth

1600000 lines in less than 10 minutes

charge\_no (id):

Primary Key format: All 1600000 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 2000000)

member\_no (id):

Foreign Key format: All 1600000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (63 : 10000)

References: member.member\_no

provider\_no (id):

Foreign Key format: All 1600000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (24 : 500)

References: provider.provider\_no

category\_no (id):

Foreign Key format: All 1600000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 10)

References: category.category\_no

charge\_dt (datetime):

Datetime format: All 1600000 datetime values are valid in the YYYYMMDD format in the range 1999/06/15 10:45 to 1999/10/13 10:51.

charge\_amt (charge):

Numerical format: All 1600000 values are numerical in the range (1.0:5001.0).

statement\_no (integer):

Numerical format: All 1600000 values are numerical in the range (0:30000).

charge\_code (code):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

NULL values (0): []

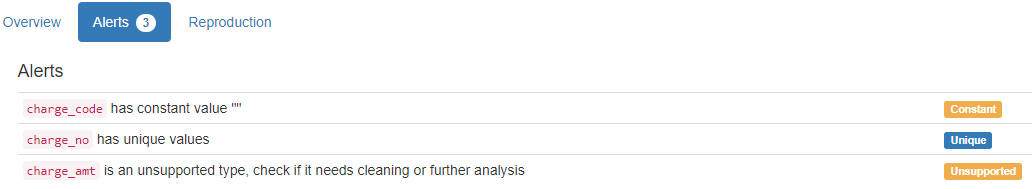
Non-NULL values (1600000): [(0, 'Empty'), (1, 'Empty'), (2, 'Empty'), (3, 'Empty'), (4, 'Empty'), (5, 'Empty'), (6, 'Empty'), (7, 'Empty'), (8, 'Empty'), (9, 'Empty'), ('...', '...'), (1599990, 'Empty'), (1599991, 'Empty'), (1599992, 'Empty'), (1599993, 'Empty'), (1599994, 'Empty'), (1599995, 'Empty'), (1599996, 'Empty'), (1599997, 'Empty'), (1599998, 'Empty'), (1599999, 'Empty')]

Categorical format with 1 unique value(s):

Category Frequency

Empty 1600000

YDATA Profiling results: It did not find the Missing values in charge\_code because they are not NULLs. They were described as Constant due to the ""(Empty) value. It also found that the PK column 'charge\_no' is Unique.



# DB.table: financial.trans – Finance – Type: Real

Over 1 million lines analysed!

trans\_id (id):

Primary Key format: All 1056320 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 3682987)

account\_id (id):

Foreign Key format: All 1056320 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 11382)

References: account.account\_id

date:

Date format: All 1056320 date values are valid in the YYYYMMDD format in the range 1993-01-01 to 1998-12-31.

type:

All 1056320 values are correctly categorical.

Categorical format with 3 unique values:

Category Frequency

VYDAJ 634571

PRIJEM 405083

VYBER 16666

operation:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

183114 Blank/Empty/Null/Missing value(s) at index(es): [(871629, NULL), (871630, NULL), (871631, NULL), (871632, NULL), (871633, NULL), (871634, NULL), (871635, NULL), (871636, NULL), (871637, NULL), (871638, NULL), ('...', '...'), (1056310, NULL), (1056311, NULL), (1056312, NULL), (1056313, NULL), (1056314, NULL), (1056315, NULL), (1056316, NULL), (1056317, NULL), (1056318, NULL), (1056319, NULL)] (displaying only the first and last 10 items)

Categorical format with 6 unique values:

Category Frequency

VYBER 434918

PREVOD NA UCET 208283

NULL 183114

VKLAD 156743

PREVOD Z UCTU 65226

VYBER KARTOU 8036

amount:

Numerical format: All 1056320 values are numerical in the range (0:87400).

balance:

Numerical format: All 1056320 values are numerical in the range (-41126:209637).

k\_symbol (symbol):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

NULL values (481881): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (871619, 'NULL'), (871620, 'NULL'), (871621, 'NULL'), (871622, 'NULL'), (871623, 'NULL'), (871624, 'NULL'), (871625, 'NULL'), (871626, 'NULL'), (871627, 'NULL'), (871628, 'NULL')]

Non-NULL values (53433): [(634, 'Blank'), (635, 'Blank'), (636, 'Blank'), (637, 'Blank'), (638, 'Blank'), (639, 'Blank'), (640, 'Blank'), (641, 'Blank'), (642, 'Blank'), (643, 'Blank'), ('...', '...'), (871323, 'Blank'), (871324, 'Blank'), (871325, 'Blank'), (871326, 'Blank'), (871327, 'Blank'), (871328, 'Blank'), (871329, 'Blank'), (871330, 'Blank'), (871331, 'Blank'), (871332, 'Blank')]

String range (lexicographical): (DUCHOD : UVER)

Frequency Distribution:

k\_symbol Frequency

NULL 481881

UROK 183114

SLUZBY 155832

SIPO 118065

Blank 53433

DUCHOD 30338

POJISTNE 18500

UVER 13580

SANKC. UROK 1577

bank:

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

782812 Blank/Empty/Null/Missing value(s) at index(es): [(0, NULL), (86, NULL), (87, NULL), (88, NULL), (89, NULL), (90, NULL), (91, NULL), (92, NULL), (93, NULL), (94, NULL), ('...', '...'), (1056310, NULL), (1056311, NULL), (1056312, NULL), (1056313, NULL), (1056314, NULL), (1056315, NULL), (1056316, NULL), (1056317, NULL), (1056318, NULL), (1056319, NULL)] (displaying only the first and last 10 items)

Frequency Distribution:

bank Frequency

NULL 782812

QR 22285

AB 21720

ST 21711

YZ 21582

GH 21499

EF 21293

KL 21234

UV 21167

OP 21094

IJ 20525

WX 20178

MN 19623

CD 19597

Range of Values: (AB to YZ)

account (integer):

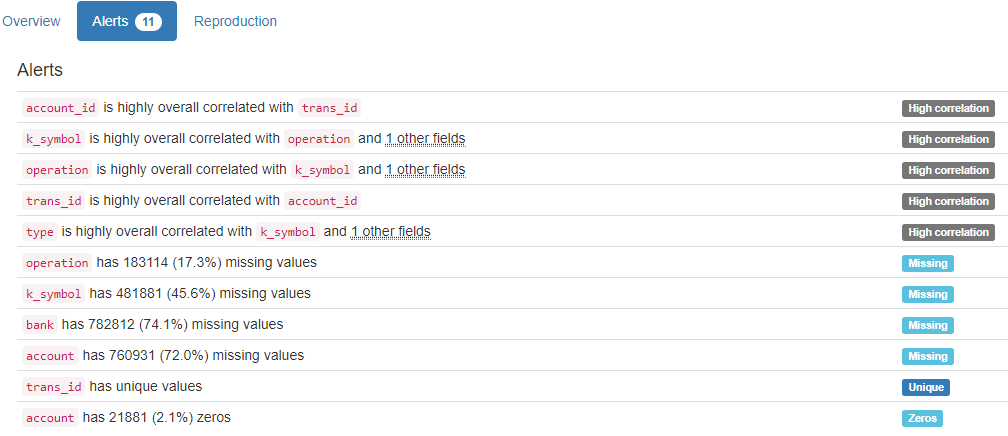
Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

760931 Blank/Empty/Null/Missing value(s) at index(es): [(0, NULL), (86, NULL), (87, NULL), (88, NULL), (89, NULL), (90, NULL), (91, NULL), (92, NULL), (93, NULL), (94, NULL), ('...', '...'), (1056310, NULL), (1056311, NULL), (1056312, NULL), (1056313, NULL), (1056314, NULL), (1056315, NULL), (1056316, NULL), (1056317, NULL), (1056318, NULL), (1056319, NULL)] (displaying only the first and last 10 items)

Range of values: (0.0:99994199.0).

YDATA Profiling results: It found 3 of the 4 Missing values columns because they contain NULLs. However, it did not find all the Missing values for column 'k\_symbol'. It just found 481881 cases of 535314 found by our system, because there are 53433 Blank values: ' '. It also found that the PK column 'trans\_id' is Unique.



# DB.table: [ftp.session](ftp://ftp.session) – Retail – Type: Synth

session\_id (id):

Primary Key format: All 30000 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (u10001 : u40000)

start\_time (datetime):

Datetime format: All 30000 datetime values are valid in the YYYYMMDD format in the range 2014/11/14 00:02 to 2014/12/22 23:58.

end\_time (datetime):

Datetime format: All 30000 datetime values are valid in the YYYYMMDD format in the range 2014/11/14 00:02 to 2014/12/22 23:58.

gender:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

15000 Blank/Empty/Null/Missing value(s) at index(es): [(15000, NULL), (15001, NULL), (15002, NULL), (15003, NULL), (15004, NULL), (15005, NULL), (15006, NULL), (15007, NULL), (15008, NULL), (15009, NULL), ('...', '...'), (29990, NULL), (29991, NULL), (29992, NULL), (29993, NULL), (29994, NULL), (29995, NULL), (29996, NULL), (29997, NULL), (29998, NULL), (29999, NULL)] (displaying only the first and last 10 items)

Categorical format with 3 unique values:

Category Frequency

NULL 15000

female 11703

male 3297

YDATA Profiling results: It found the same 1 Missing values column because they contain NULLs. It also found that the PK column 'session\_id' is Unique.

A white background with black lines

Description automatically generated

# DB.table: Hepatitis\_std.dispat – Medicine – Type: Real

m\_id (id):

Primary Key format: All 500 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 500)

sex:

All 500 values are correctly categorical.

Categorical format with 2 unique values:

Category Frequency

0 355

1 145

age:

Age format: All 500 values are numerical and valid in the range [0, 130].

Actual range of values: (0 : 6)

Type:

All 500 values are correctly categorical.

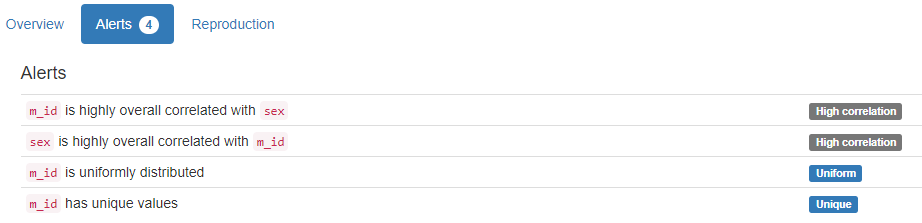
Categorical format with 2 unique values:

Category Frequency

1 294

1. 206

YDATA Profiling results: It just found out that the PK 'm\_id' has unique values and is uniformly distributed:



# DB.table: Hockey.Master – Sport – Type: Real

playerID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

241 Blank/Empty/Null/Missing value(s) at index(es): [(22, 'NULL'), (31, 'NULL'), (38, 'NULL'), (40, 'NULL'), (41, 'NULL'), (66, 'NULL'), (152, 'NULL'), (172, 'NULL'), (214, 'NULL'), (243, 'NULL'), ('...', '...'), (7447, 'NULL'), (7456, 'NULL'), (7499, 'NULL'), (7597, 'NULL'), (7615, 'NULL'), (7616, 'NULL'), (7680, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7739, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (aaltoan01 : zyuzian01)

coachID (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

7366 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Alphanumeric range of values: (abelsi01c : youngwe02c)

References: Coaches.coachID

hofID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):   
 7395 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (abelsi01h : ziegljo01h)

firstName (name):

Name format: Error(s) found

DQI #1 (Missing Data - Completeness):

13 Blank/Empty/Null/Missing value(s) at index(es): [(407, 'NULL'), (779, 'NULL'), (790, 'NULL'), (1102, 'NULL'), (2582, 'NULL'), (2598, 'NULL'), (2717, 'NULL'), (3526, 'NULL'), (3879, 'NULL'), (4311, 'NULL'), (4346, 'NULL'), (5687, 'NULL'), (6106, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

1 Capitalization/Format issue(s) at index(es): [(1357, "D'arcy")]

Frequency Distribution:

firstName Frequency

Mike 193

Bob 142

John 136

Bill 130

Jim 114

Dave 105

Steve 103

Brian 84

Paul 83

Chris 79

... ...

Yanic 1

Yevgeny 1

Yutaka 1

Zarley 1

Zbynek 1

Zellio 1

Zenith 1

Zenon 1

Ziggy 1

Zinetul 1

Range of Values: (Aaron to Zinetul)

lastName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

46 Capitalization/Format issue(s) at index(es): [(1569, 'DeBlois'), (1570, 'DeBoer'), (1572, 'DeBrusk'), (1575, 'DeFauw'), (1578, 'DeGray'), (1581, 'DeJordy'), (1583, 'DelGuidice'), (1597, 'DeMarco'), (1598, 'DeMarco'), (1612, 'DePalma'), ('...', '...'), (3957, 'LeBrun'), (3960, 'LeClair'), (3961, 'LeClair'), (3972, 'LeDuc'), (4029, 'LeNeveu'), (4045, 'LeSueur'), (4152, 'LoPresti'), (4153, 'LoPresti'), (7267, 'Van der Gulik'), (7270, 'VandenBussche')] (displaying only the first and last 10 items)

Frequency Distribution:

lastName Frequency

Smith 69

Brown 32

Wilson 31

Johnson 27

Murray 26

Stewart 22

Miller 20

Anderson 19

Martin 17

Jackson 14

... ...

Zubarev 1

Zubrus 1

Zuccarello 1

Zucker 1

Zuk 1

Zuke 1

Zunich 1

Zyuzin 1

de Haan 1

de Vries 1

Range of Values: (Aalto to de Vries)

nameNote (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

7743 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution:

nameNote Frequency

NULL 7743

also known as Bellehumeur 1

also known as Block 1

also known as Bowcher 1

also known as Bremberg 1

also known as Burmeister 1

also known as Coleman 1

also known as Couture 1

also known as Desrivieres 1

also known as Kahibaitche 1

also known as Kessler 1

also known as Landiak 1

also known as Linton Muldoon Tracey 1

also known as Mike Jefferson 1

also known as Wochy 1

also listed as Bourdginon 1

born Bodnarchuk 1

born Guoth 1

born Tselios 1

Range of Values: (also known as Bellehumeur to born Tselios)

nameGiven (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1776 Blank/Empty/Null/Missing value(s) at index(es): [(4, 'NULL'), (9, 'NULL'), (17, 'NULL'), (18, 'NULL'), (23, 'NULL'), (34, 'NULL'), (35, 'NULL'), (38, 'NULL'), (45, 'NULL'), (62, 'NULL'), ('...', '...'), (7740, 'NULL'), (7741, 'NULL'), (7743, 'NULL'), (7744, 'NULL'), (7745, 'NULL'), (7748, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7755, 'NULL'), (7756, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

9 Capitalization/Format issue(s) at index(es): [(805, 'Neal LaMoy'), (1357, "D'arcy John"), (2507, 'Harrison LeRoy'), (2540, 'MIchael James'), (3290, 'Edward LaVern'), (3687, 'JArda'), (3992, 'CLaude'), (4420, 'Albert LeRoy'), (5909, 'Robert DeHart')]

Frequency Distribution:

nameGiven Frequency

NULL 1776

Michael 62

Robert 49

John 43

James 36

Daniel 35

David 35

Peter 31

Steven 26

William 26

Paul 25

... ...

Yvan Joseph Rene 1

Yvan Serge 1

Yvon Jean 1

Yvon Jules 1

Yvon Pierre 1

Zarley Bennett 1

Zdeno 1

Zellio Louis Peter 1

Zenith 1

Zigmund 1

Range of Values: (A.B. to Zigmund)

nameNick (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

6455 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (13, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

1 Extraneous data value(s) at index(es): [(1364, '7-11')]

DQI #15 (Domain Violation - Accuracy):

6 Capitalization/Format issue(s) at index(es): [(307, 'Goal-a-Game'), (1942, 'X-et'), (2004, 'FedEx'), (2202, 'CoCo'), (5089, 'Sweet Lou from the Soo'), (7043, 'Larry/Half-n-Half')]

Frequency Distribution:

nameNick Frequency

NULL 6455

Red 24

Moose 13

Bill 7

Chief 7

Doc 7

Killer 7

Rocky 7

Duke 6

Bam Bam 5

Cowboy 5

... ...

Wolfman 1

Woodsy 1

Wrecker/Herc 1

X-et 1

Yank 1

Yug 1

Zabrda 1

Zacky 1

Zeus 1

Zholi 1

Range of Values: (7-11 to Zholi)

height:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

427 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (22, 'NULL'), (31, 'NULL'), (38, 'NULL'), (40, 'NULL'), (41, 'NULL'), (45, 'NULL'), (66, 'NULL'), (116, 'NULL'), ('...', '...'), (7502, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7597, 'NULL'), (7615, 'NULL'), (7616, 'NULL'), (7680, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7739, 'NULL')] (displaying only the first and last 10 items)

Range of values: (63:81).

weight:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):   
 425 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (22, 'NULL'), (31, 'NULL'), (38, 'NULL'), (40, 'NULL'), (41, 'NULL'), (45, 'NULL'), (66, 'NULL'), (116, 'NULL'), ('...', '...'), (7502, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7597, 'NULL'), (7615, 'NULL'), (7616, 'NULL'), (7680, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7739, 'NULL')] (displaying only the first and last 10 items)

Range of values: (125:265).

shootCatch (category):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

713 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (18, 'NULL'), (22, 'NULL'), (31, 'NULL'), (38, 'NULL'), (40, 'NULL'), (41, 'NULL'), (45, 'NULL'), (64, 'NULL'), (66, 'NULL'), ('...', '...'), (7680, 'NULL'), (7685, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7737, 'NULL'), (7739, 'NULL'), (7748, 'NULL'), (7751, 'NULL'), (7755, 'NULL'), (7756, 'NULL')]

Categorical format with 4 unique value(s):

Category Frequency

L 4604

R 2443

NULL 713

B 1

legendsID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1184 Blank/Empty/Null/Missing value(s) at index(es): [(1, 'NULL'), (10, 'NULL'), (15, 'NULL'), (17, 'NULL'), (18, 'NULL'), (20, 'NULL'), (32, 'NULL'), (38, 'NULL'), (42, 'NULL'), (45, 'NULL'), ('...', '...'), (7713, 'NULL'), (7733, 'NULL'), (7737, 'NULL'), (7742, 'NULL'), (7748, 'NULL'), (7750, 'NULL'), (7751, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7758, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (10002 : P200904)

ihdbID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

636 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (18, 'NULL'), (22, 'NULL'), (31, 'NULL'), (40, 'NULL'), (41, 'NULL'), (64, 'NULL'), (66, 'NULL'), (100, 'NULL'), (116, 'NULL'), ('...', '...'), (7616, 'NULL'), (7680, 'NULL'), (7685, 'NULL'), (7702, 'NULL'), (7737, 'NULL'), (7739, 'NULL'), (7748, 'NULL'), (7751, 'NULL'), (7755, 'NULL'), (7756, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (100 : 9999)

hrefID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):   
 304 Blank/Empty/Null/Missing value(s) at index(es): [(20, 'NULL'), (22, 'NULL'), (31, 'NULL'), (40, 'NULL'), (41, 'NULL'), (66, 'NULL'), (94, 'NULL'), (99, 'NULL'), (152, 'NULL'), (167, 'NULL'), ('...', '...'), (7500, 'NULL'), (7502, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7585, 'NULL'), (7606, 'NULL'), (7615, 'NULL'), (7616, 'NULL'), (7702, 'NULL'), (7739, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (aaltoan01 : zyuzian01)

firstNHL (first):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

910 Blank/Empty/Null/Missing value(s) at index(es): [(1, 'NULL'), (10, 'NULL'), (15, 'NULL'), (17, 'NULL'), (20, 'NULL'), (22, 'NULL'), (31, 'NULL'), (32, 'NULL'), (38, 'NULL'), (40, 'NULL'), ('...', '...'), (7640, 'NULL'), (7668, 'NULL'), (7680, 'NULL'), (7689, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7739, 'NULL'), (7742, 'NULL'), (7750, 'NULL'), (7758, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (1917 : 2011)

lastNHL (last):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

910 Blank/Empty/Null/Missing value(s) at index(es): [(1, 'NULL'), (10, 'NULL'), (15, 'NULL'), (17, 'NULL'), (20, 'NULL'), (22, 'NULL'), (31, 'NULL'), (32, 'NULL'), (38, 'NULL'), (40, 'NULL'), ('...', '...'), (7640, 'NULL'), (7668, 'NULL'), (7680, 'NULL'), (7689, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7739, 'NULL'), (7742, 'NULL'), (7750, 'NULL'), (7758, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (1917 : 2011)

firstWHA (first):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

6858 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (9, 'NULL'), (12, 'NULL'), (13, 'NULL'), ('...', '...'), (7748, 'NULL'), (7749, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (1972 : 1978)

lastWHA (last):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

6858 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (9, 'NULL'), (12, 'NULL'), (13, 'NULL'), ('...', '...'), (7748, 'NULL'), (7749, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (1972 : 1978)

pos (position):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

314 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (22, 'NULL'), (31, 'NULL'), (38, 'NULL'), (40, 'NULL'), (41, 'NULL'), (45, 'NULL'), (66, 'NULL'), (116, 'NULL'), (152, 'NULL'), ('...', '...'), (7378, 'NULL'), (7456, 'NULL'), (7500, 'NULL'), (7574, 'NULL'), (7615, 'NULL'), (7616, 'NULL'), (7680, 'NULL'), (7695, 'NULL'), (7702, 'NULL'), (7739, 'NULL')]

Categorical format with 19 unique value(s):

Category Frequency

D 2218

C 1605

L 1450

R 1348

G 761

NULL 314

F 43

R/C 5

C/D 3

C/L 2

C/R 2

D/L 2

L/C 2

D/C 1

D/R 1

L/D 1

R/D 1

R/L 1

W 1

birthYear (year):

Year format: Error(s) found:

DQI #1 (Missing Data - Completeness):

218 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (45, 'NULL'), (116, 'NULL'), (211, 'NULL'), (246, 'NULL'), (248, 'NULL'), (352, 'NULL'), (389, 'NULL'), (391, 'NULL'), ('...', '...'), (7328, 'NULL'), (7334, 'NULL'), (7347, 'NULL'), (7348, 'NULL'), (7375, 'NULL'), (7502, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7702, 'NULL'), (7747, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (1841 : 1993)

birthMon (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

255 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (45, 'NULL'), (116, 'NULL'), (211, 'NULL'), (246, 'NULL'), (248, 'NULL'), (282, 'NULL'), (352, 'NULL'), (389, 'NULL'), ('...', '...'), (7348, 'NULL'), (7375, 'NULL'), (7378, 'NULL'), (7403, 'NULL'), (7502, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7606, 'NULL'), (7702, 'NULL'), (7747, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (1 : 9)

birthDay (day):

Day format: Error(s) found:

DQI #1 (Missing Data - Completeness):   
 257 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (45, 'NULL'), (116, 'NULL'), (211, 'NULL'), (246, 'NULL'), (248, 'NULL'), (282, 'NULL'), (352, 'NULL'), (389, 'NULL'), ('...', '...'), (7348, 'NULL'), (7375, 'NULL'), (7378, 'NULL'), (7403, 'NULL'), (7502, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7606, 'NULL'), (7702, 'NULL'), (7747, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (1 : 31)

birthCountry (country):

Country format: Error(s) found:

DQI #1 (Missing Data - Completeness):

365 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (45, 'NULL'), (64, 'NULL'), (100, 'NULL'), (116, 'NULL'), (124, 'NULL'), (174, 'NULL'), (186, 'NULL'), (211, 'NULL'), ('...', '...'), (7485, 'NULL'), (7502, 'NULL'), (7527, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7685, 'NULL'), (7702, 'NULL'), (7737, 'NULL'), (7748, 'NULL'), (7756, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 50 countries):

Country Frequency

Canada 5328

USA 994

NULL 365

Sweden 233

Czech Republic 206

Russia 172

Finland 164

Slovakia 64

Germany 28

England 25

... ...

Netherlands 1

Nigeria 1

Northern Ireland 1

Paraguay 1

Slovenia 1

South Africa 1

South Korea 1

South Wales 1

Taiwan 1

Tanzania 1

birthState (state):

State format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1437 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (10, 'NULL'), (11, 'NULL'), (17, 'NULL'), (20, 'NULL'), (23, 'NULL'), (34, 'NULL'), (35, 'NULL'), (37, 'NULL'), (45, 'NULL'), ('...', '...'), (7744, 'NULL'), (7745, 'NULL'), (7748, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 58 states/provinces):

State Frequency

ON 2365

NULL 1437

QC 898

AB 570

SK 527

MB 414

BC 357

MN 240

MA 186

MI 140

... ...

County Wexford 1

DE 1

IA 1

KS 1

MT 1

NC 1

NE 1

WY 1

YT 1

Yorkshire 1

birthCity (city):

City format: Error(s) found:

DQI #1 (Missing Data - Completeness):

373 Blank/Empty/Null/Missing value(s) at index(es): [(17, 'NULL'), (20, 'NULL'), (40, 'NULL'), (45, 'NULL'), (64, 'NULL'), (100, 'NULL'), (116, 'NULL'), (124, 'NULL'), (174, 'NULL'), (186, 'NULL'), ('...', '...'), (7485, 'NULL'), (7502, 'NULL'), (7527, 'NULL'), (7574, 'NULL'), (7579, 'NULL'), (7685, 'NULL'), (7702, 'NULL'), (7737, 'NULL'), (7748, 'NULL'), (7756, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

2 Extraneous data value(s) at index(es): [(4886, '100 Mile House'), (6021, '100 Mile House')]

DQI #15 (Domain Violation - Accuracy):

44 Capitalization/Format issue(s) at index(es): [(168, 'Portage la Prairie'), (192, "Val d'Or"), (193, 'Cap-de-la-Madeleine'), (354, 'Pointe-aux-Trembles'), (399, 'Portage la Prairie'), (401, 'Portage la Prairie'), (648, "Val d'Or"), (653, 'Notre-Dame-du-Lac'), (866, 'Usti nad Labem'), (976, 'Usti nad Labem'), ('...', '...'), (5743, 'Lac la Tortue'), (5851, 'LaSalle'), (6031, 'Cap-de-la-Madeleine'), (6175, "Val d'Or"), (6231, 'St.-Gabriel-de-Bran.'), (6793, 'LaSalle'), (6964, 'Cap del la Madeleine'), (6989, 'Saint-Methode-de-Fro'), (7161, 'Pointe-aux-Trembles'), (7504, 'WInnipeg')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 1914 cities):

City Frequency

Toronto 439

NULL 373

Montreal 322

Winnipeg 203

Edmonton 186

Ottawa 173

Calgary 117

Regina 92

Sudbury 75

Saskatoon 72

... ...

Yonkers 1

Young 1

Youngstown 1

Zabrize 1

Zaria 1

Zilina 1

Ziln 1

Zranjanin 1

Zug 1

Zweibrucken 1

deathYear (year):

Year format: Error(s) found:

DQI #1 (Missing Data - Completeness):

6441 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (7750, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

93 Value(s) outside range [1800, 2100] at index(es): [(88, '9999'), (397, '9999'), (430, '9999'), (674, '9999'), (750, '9999'), (792, '9999'), (822, '9999'), (897, '9999'), (901, '9999'), (916, '9999'), ('...', '...'), (6701, '9999'), (6743, '9999'), (6758, '9999'), (6799, '9999'), (6864, '9999'), (6909, '9999'), (7057, '9999'), (7060, '9999'), (7159, '9999'), (7606, '9999')] (displaying only the first and last 10 items)

Actual range of values: (1907 : 9999)

deathMon (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

6588 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (7750, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (1 : 9)

deathDay (day):

Day format: Error(s) found:

DQI #1 (Missing Data - Completeness):

6593 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (7750, 'NULL'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (1 : 31)

deathCountry (country):

Country format: Error(s) found:

DQI #1 (Missing Data - Completeness):

7263 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution:

Country Frequency

NULL 7263

Canada 343

USA 128

Russia 11

Czech Republic 3

Germany 2

Sweden 2

Switzerland 2

Belarus 1

Belgium 1

France 1

Holland 1

Italy 1

Norway 1

Turkey 1

deathState (state):

State format: Error(s) found:   
 7290 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 45 states/provinces):

State Frequency

NULL 7290

ON 197

QC 40

BC 29

AB 26

MB 24

SK 18

MN 15

FL 14

CA 12

... ...

MS 1

NC 1

ND 1

NL 1

NT 1

NV 1

OR 1

PE 1

RI 1

TN 1

deathCity (city):

City format: Error(s) found:

DQI #1 (Missing Data - Completeness):

7278 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (7751, 'NULL'), (7752, 'NULL'), (7753, 'NULL'), (7754, 'NULL'), (7755, 'NULL'), (7756, 'NULL'), (7757, 'NULL'), (7758, 'NULL'), (7759, 'NULL'), (7760, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

2 Capitalization/Format issue(s) at index(es): [(1004, 'Salaberry-de-Valleyfield'), (5616, 'Ile du Pas')]

Frequency Distribution (showing top and bottom 10 of 240 cities):

City Frequency

NULL 7278

Toronto 45

Ottawa 30

Winnipeg 19

Montreal 16

Calgary 11

Vancouver 10

Yaroslavl 9

Sudbury 8

Edmonton 7

... ...

Vernon 1

Victoriaville 1

Viking 1

Westfield 1

Westlake 1

Weyburn 1

White Rock 1

Wyandotte 1

Yellowknife 1

Zurich 1

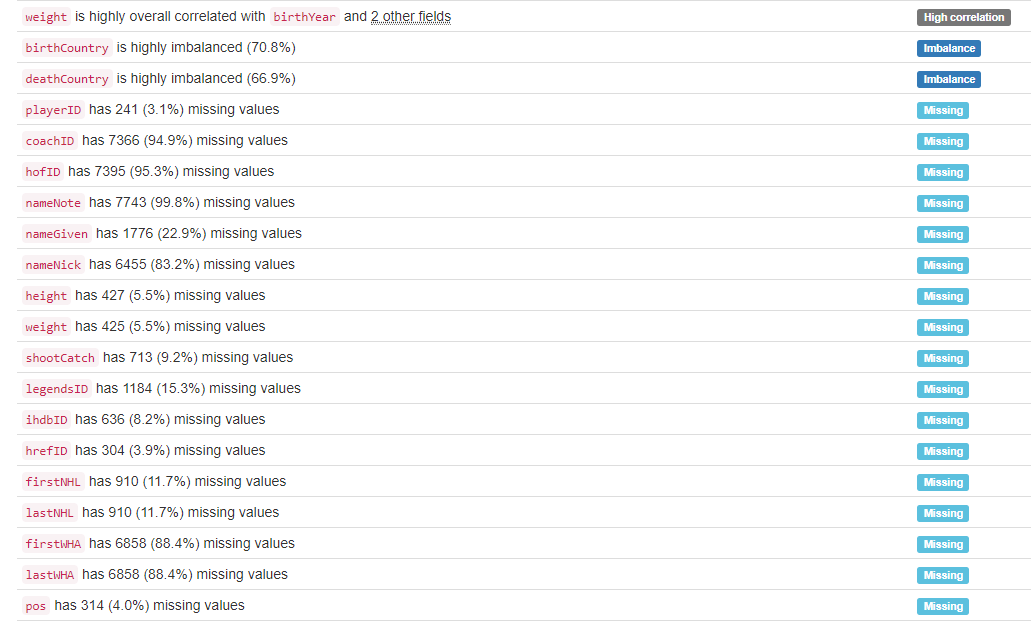
YDATA Profiling results: It found the same 30 Missing values columns because they contain NULLs. This table has no Primary Key, just one Foreign Key defined.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



A close-up of a computer screen

Description automatically generated

# DB.table: imdb\_full.actors – Entertainment – Type: Real

actorid (id):

Primary Key format: All 98690 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (4 : 2749386)

name:

Name format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

14 Extraneous data value(s) at index(es): [(4452, 'Berg, Dav!d'), (5058, 'Bird No. 9'), (15039, "Demitro, Michael '2-Smoove'"), (17032, 'E-40'), (18209, 'Fab 5 Freddy'), (36455, 'Mack 10'), (42646, 'N!xau'), (43661, "No's. 1 to 84"), (43748, 'Nooney Rickett 4, The'), (60264, 'UB40'), (65794, '7 Year Bitch'), (65795, '7, Lucky'), (76160, 'Fuschia!'), (98305, 'Y?N-Vee')]

DQI #15 (Domain Violation - Accuracy):

692 Capitalization/Format issue(s) at index(es): [(0, '$hort, Too'), (103, "Abossolo M'bo, Emile"), (529, 'al-Assad, Hafez'), (886, 'Allen, LeRoy'), (2230, 'AuCoin, Kelly'), (2921, "Banai, Ya'ackov"), (5808, "Bonds, De'aundre"), (5892, 'Bookwalter, DeVeren'), (8124, 'Burton, LeVar'), (8395, 'Bôle du Chaumont, Gérard'), ('...', '...'), (96632, 'von der Heyde, Nicole'), (96968, 'Ward, LaToya'), (97668, 'Willette, JoAnn'), (97709, 'Williams, JoBeth'), (97754, "Williams, Swa'vet"), (97787, 'Willis, JoAnne C.'), (97796, 'Willis, Scout LaRue'), (97873, 'Wilson, PaSean'), (98154, "Woolfolk, A'da Alison"), (98240, 'Wu, Chien-lien')] (displaying only the first and last 10 items)

Frequency Distribution:

name Frequency

Deroche, Craig 2

DuBois, Diane 2

McKinney, Bill (I) 2

de la Torre, Antonio (I) 2

$hort, Too 1

'Bear'Boyd, Steven 1

'Dead End' Kids, The 1

7 Year Bitch 1

7, Lucky 1

Aaby, Kristian 1

... ...

Ôtsuka, Akio 1

Ôtsuka, Kunio 1

Ôtsuki, Shûji 1

Öhrstedt, Pekka 1

Öldutúnsskóla, Kór 1

Örnólfsdóttir, Álfrún 1

Öss, Eniko 1

Øby, Kalle 1

Øksnes, Erik 1

Überzone 1

Range of Values: ($hort, Too to Überzone)

sex:

Categorical format: All 98690 values are correctly categorical.

Categorical format with 2 unique value(s):

Category Frequency

M 65794

F 32896

YDATA Profiling results: It just found out that the PK 'actorid' has unique values:

A white rectangular object with a black line

Description automatically generated

# DB.table: imdb\_full.movies – Entertainment – Type: Real

movieid (id):

Primary Key format: All 3832 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1672052 : 2593313)

title:

Categorical format: Error(s) found:

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 3832 unique value(s):

Category Frequency

'Breaker' Morant (1980) 1

'Crocodile' Dundee II (1988) 1

'Til There Was You (1997) 1

'night, Mother (1986) 1

...And Justice for All (2014) 1

1-2-3-4 ou Les collants noirs (1961) 1

1-900 (1996) 1

10 Things I Hate About You (1999) 1

101 Dalmatians (1996) 1

12 Angry Men (1957) 1

... ...

Zbogum na dvaesetiot vek (1998) 1

Zerkalo (1975) 1

Zero Effect (1998) 1

Zeus and Roxanne (1997) 1

Zone 39 (1996) 1

eXistenZ (1999) 1

¡Three Amigos! (1986) 1

À nous la liberté (1931) 1

Á köldum klaka (1995) 1

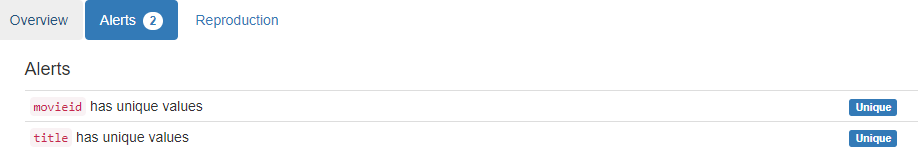
Átame! (1990) 1

year:

Year format: All 3832 values are numerical and valid in the range [1800, 2100].

Actual range of values: (1898 : 2014)

YDATA Profiling results: It just found out that the PK column 'movieid' has unique values:



# DB.table: imdb\_full.movies2actors – Entertainment – Type: Real

movieid|actorid (Composite Primary Key):

Composite Primary Key format: All 138349 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (1672052|1045133 : 2593313|947005)

as\_character (char):

String format: Error(s) found:

NULL values (0): []

Non-NULL values (857): [(312, 'Empty'), (339, 'Empty'), (1200, 'Empty'), (2283, 'Empty'), (3081, 'Empty'), (4488, 'Empty'), (4994, 'Empty'), (5412, 'Empty'), (5798, 'Empty'), (8141, 'Empty'), ('...', '...'), (137963, 'Empty'), (137964, 'Empty'), (137965, 'Empty'), (137971, 'Empty'), (137973, 'Empty'), (137976, 'Empty'), (138083, 'Empty'), (138085, 'Empty'), (138089, 'Empty'), (138096, 'Empty')]

String range (lexicographical): ((also archive footage) [Babe' Hanson] <5> : [úttörõtanár] <6>)

Frequency Distribution:

as\_character Frequency

Empty 857

[Himself] 336

(archive footage) (uncredited) [Himself] 219

(archive footage) [Himself] 189

(archive footage) 134

[Herself] 102

[Extra] 67

<20> 48

<23> 47

<26> 47

... ...

[Élisabeth] <3> 1

[Émile] <1> 1

[Émilia] <7> 1

[Émilie Savignard] <5> 1

[Éponine] <27> 1

[Étienne Lantier] <2> 1

[Étienne, adulte] <14> 1

[Étienne] <4> 1

[Ügyelõ] <15> 1

[úttörõtanár] <6> 1

leading (lead):

Numerical format: All 138349 values are numerical in the range (0:307).

YDATA Profiling results: It just found out that the composite PK movieid|actorid has unique values. It did not find the 857 Empty cases in the as\_character column:



# DB.table: imdb\_MovieLens.users – Entertainment – Type: Real

userid (id):

Primary Key format: All 6039 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 6040)

age:

Age format: All 6039 values are numerical and valid in the range [0, 130].

Actual range of values: (1 : 56)

u\_gender (gender):

Categorical format: All 6039 values are correctly categorical.

Categorical format with 2 unique value(s):

Category Frequency

M 4331

F 1708

occupation:

Categorical format: All 6039 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

2 1614

5 1453

1 1132

4 1051

3 789

YDATA Profiling results: It just found out that the PK column 'userid' has unique values and is uniformly distributed:: A white background with black lines

Description automatically generated

# DB.table: imdb\_MovieLens.movies – Entertainment – Type: Real

movieid (id):

Primary Key format: All 3832 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1672052 : 2593313)

year:

Year format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

3832 Value(s) outside range [1800, 2100] at index(es): [(0, 3), (1, 4), (2, 4), (3, 4), (4, 4), (5, 3), (6, 4), (7, 4), (8, 4), (9, 3), ('...', '...'), (3822, 4), (3823, 3), (3824, 4), (3825, 4), (3826, 4), (3827, 4), (3828, 4), (3829, 2), (3830, 4), (3831, 4)] (displaying only the first and last 10 items)

Actual range of values: (1 : 4)

isEnglish (enum):

Categorical format: All 3832 values are correctly categorical.

Categorical format with 2 unique value(s):

Category Frequency

T 3411

F 421

country:

Country format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

531 Capitalization/Format issue(s) at index(es): [(0, 'other'), (1, 'other'), (21, 'other'), (40, 'other'), (51, 'other'), (61, 'other'), (65, 'other'), (67, 'other'), (96, 'other'), (106, 'other'), ('...', '...')]

Frequency Distribution:

Country Frequency

USA 2725

other 531

UK 384

France 192

runningtime (integer):

Numerical format: All 3832 values are numerical in the range (0:3).

YDATA Profiling results: It just found out that the PK column 'movieid' has unique values:

A white background with black lines

Description automatically generated

# DB.table: imdb\_MovieLens.u2base – Entertainment – Type: Real

Almost 1 million lines

userid|movieid (Composite Primary Key):

Composite Primary Key format: All 996159 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (1000|1674388 : 9|2577314)

rating:

Categorical format: All 996159 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

4 348150

3 260180

5 225108

2 107114

1 55607

YDATA Profiling results: It just found out that the Composite PK userid|movieid has unique values:

A white background with black and white clouds

Description automatically generated

# DB.table: lahman\_2014.halloffame – Sport – Type: Real

hofID|yearID (Composite Primary Key):

Composite Primary Key format: All 3883 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (aaronha01h|1982 : ziskri01h|1989)

votedBy (varchar):

String format: All 3883 string values are valid.

String range (lexicographical): (BBWAA : Veterans)

ballots (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

169 Blank/Empty/Null/Missing value(s) at index(es): [(51, 'NULL'), (61, 'NULL'), (63, 'NULL'), (96, 'NULL'), (105, 'NULL'), (130, 'NULL'), (146, 'NULL'), (150, 'NULL'), (153, 'NULL'), (173, 'NULL'), ('...', '...'), (3727, 'NULL'), (3749, 'NULL'), (3750, 'NULL'), (3757, 'NULL'), (3789, 'NULL'), (3802, 'NULL'), (3827, 'NULL'), (3840, 'NULL'), (3848, 'NULL'), (3874, 'NULL')] (displaying only the first and last 10 items)

Range of values: (78.0:581.0).

needed (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

251 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (51, 'NULL'), (61, 'NULL'), (63, 'NULL'), (77, 'NULL'), (96, 'NULL'), (105, 'NULL'), (125, 'NULL'), (130, 'NULL'), (134, 'NULL'), ('...', '...'), (3750, 'NULL'), (3757, 'NULL'), (3789, 'NULL'), (3802, 'NULL'), (3814, 'NULL'), (3827, 'NULL'), (3840, 'NULL'), (3848, 'NULL'), (3863, 'NULL'), (3874, 'NULL')] (displaying only the first and last 10 items)

Range of values: (59.0:436.0).

votes (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

169 Blank/Empty/Null/Missing value(s) at index(es): [(51, 'NULL'), (61, 'NULL'), (63, 'NULL'), (96, 'NULL'), (105, 'NULL'), (130, 'NULL'), (146, 'NULL'), (150, 'NULL'), (153, 'NULL'), (173, 'NULL'), ('...', '...'), (3727, 'NULL'), (3749, 'NULL'), (3750, 'NULL'), (3757, 'NULL'), (3789, 'NULL'), (3802, 'NULL'), (3827, 'NULL'), (3840, 'NULL'), (3848, 'NULL'), (3874, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:537.0).

inducted (varchar):

String format: All 3883 string values are valid.

String range (lexicographical): (N : Y)

category:

Categorical format: All 3883 values are correctly categorical.

Categorical format with 4 unique value(s):

Category Frequency

Player 3765

Manager 69

Pioneer/Executive 39

Umpire 10

YDATA Profiling results: It found the same 3 Missing values columns because they contain NULLs. It also found that the composite PK column 'hof\_id|yearID' is Unique.

A screenshot of a computer

Description automatically generated

# DB.table: lahman\_2014.players – Sport – Type: Real

lahmanID (id):

Primary Key format: All 16407 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 17610)

playerID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2 Duplicate value(s) at index(es): [[456, 'baezda01'], [457, 'baezda01']]

DQI #19 (Uniqueness Violation - Uniqueness):

1 Uniqueness violation(s) at index(es): [(457, 'baezda01')]

Alphanumeric range of values: (aardsda01 : zwilldu01)

managerID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

15850 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (adcocjo01m : zimmedo01m)

hofID (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

15287 Blank/Empty/Null/Missing value(s) at index(es): [(1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Alphanumeric range of values: (aaronha01h : ziskri01h)

References: halloffame.hofID

birthYear (year):

Year format: Error(s) found:

DQI #1 (Missing Data - Completeness):

230 Blank/Empty/Null/Missing value(s) at index(es): [(56, 'NULL'), (95, 'NULL'), (179, 'NULL'), (556, 'NULL'), (617, 'NULL'), (653, 'NULL'), (654, 'NULL'), (972, 'NULL'), (1018, 'NULL'), (1203, 'NULL'), ('...', '...'), (14931, 'NULL'), (15024, 'NULL'), (15119, 'NULL'), (15141, 'NULL'), (15313, 'NULL'), (15383, 'NULL'), (15476, 'NULL'), (15550, 'NULL'), (15567, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (1831.0 : 1986.0)

birthMonth (month):

Month format: Error(s) found:

DQI #1 (Missing Data - Completeness):

445 Blank/Empty/Null/Missing value(s) at index(es): [(56, 'NULL'), (59, 'NULL'), (95, 'NULL'), (113, 'NULL'), (177, 'NULL'), (179, 'NULL'), (345, 'NULL'), (392, 'NULL'), (465, 'NULL'), (507, 'NULL'), ('...', '...'), (15401, 'NULL'), (15476, 'NULL'), (15514, 'NULL'), (15550, 'NULL'), (15564, 'NULL'), (15567, 'NULL'), (15801, 'NULL'), (15988, 'NULL'), (16198, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution:

Month Frequency

August 1552

October 1508

November 1399

January 1391

September 1374

December 1335

March 1315

July 1295

May 1218

April 1209

February 1204

June 1162

birthDay (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

585 Blank/Empty/Null/Missing value(s) at index(es): [(56, 'NULL'), (59, 'NULL'), (76, 'NULL'), (95, 'NULL'), (113, 'NULL'), (177, 'NULL'), (179, 'NULL'), (345, 'NULL'), (367, 'NULL'), (392, 'NULL'), ('...', '...'), (15476, 'NULL'), (15514, 'NULL'), (15550, 'NULL'), (15564, 'NULL'), (15567, 'NULL'), (15801, 'NULL'), (15805, 'NULL'), (15988, 'NULL'), (16198, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:31.0).

birthCountry (country):

Country format: Error(s) found:

DQI #1 (Missing Data - Completeness):

278 Blank/Empty/Null/Missing value(s) at index(es): [(56, 'NULL'), (95, 'NULL'), (556, 'NULL'), (617, 'NULL'), (653, 'NULL'), (654, 'NULL'), (731, 'NULL'), (972, 'NULL'), (1018, 'NULL'), (1203, 'NULL'), ('...', '...'), (14931, 'NULL'), (15024, 'NULL'), (15119, 'NULL'), (15141, 'NULL'), (15313, 'NULL'), (15383, 'NULL'), (15476, 'NULL'), (15550, 'NULL'), (15567, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

1 Capitalization/Format issue(s) at index(es): [(11255, 'A Ship on Atlantic Ocean')]

Frequency Distribution (showing top and bottom 10 of 54 countries):

Country Frequency

USA 14524

D.R. 411

NULL 278

P.R. 215

CAN 205

Venezuela 181

Cuba 150

Mexico 98

Panama 45

Ireland 40

... ...

China 1

Denmark 1

Finland 1

Greece 1

Honduras 1

Okinawa 1

Philippines 1

Singapore 1

South Vietnam 1

Switzerland 1

birthState (state):

State format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1578 Blank/Empty/Null/Missing value(s) at index(es): [(30, 'NULL'), (33, 'NULL'), (39, 'NULL'), (40, 'NULL'), (41, 'NULL'), (56, 'NULL'), (91, 'NULL'), (92, 'NULL'), (95, 'NULL'), (99, 'NULL'), ('...', '...'), (16396, 'NULL'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

98 Extraneous data value(s) at index(es): [(34, '6'), (38, '25'), (186, '26'), (216, '26'), (222, '26'), (229, '30'), (424, '30'), (644, '30'), (652, '30'), (667, '26'), ('...', '...'), (15632, '25'), (15633, '26'), (15907, '25'), (15908, '19'), (15909, '26'), (15910, '19'), (15911, '19'), (16158, '19'), (16359, '2'), (16361, '2')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 80 states/provinces):

State Frequency

CA 1793

NULL 1578

PA 1320

NY 1105

IL 976

OH 952

TX 702

MA 635

MO 549

MI 403

... ...

14 2

15 2

20 2

28 2

7 2

PE 2

16 1

21 1

NF 1

PH 1

birthCity (city):

City format: Error(s) found:

DQI #1 (Missing Data - Completeness):

392 Blank/Empty/Null/Missing value(s) at index(es): [(56, 'NULL'), (95, 'NULL'), (99, 'NULL'), (478, 'NULL'), (544, 'NULL'), (556, 'NULL'), (613, 'NULL'), (617, 'NULL'), (653, 'NULL'), (654, 'NULL'), ('...', '...'), (15313, 'NULL'), (15345, 'NULL'), (15383, 'NULL'), (15476, 'NULL'), (15550, 'NULL'), (15567, 'NULL'), (15805, 'NULL'), (15988, 'NULL'), (16198, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

119 Capitalization/Format issue(s) at index(es): [(158, 'DeLand'), (187, 'ElPaso'), (311, 'ElTigre'), (356, 'ElPaso'), (813, 'ElDorado'), (842, 'ElDorado'), (961, 'LaSalle'), (1147, 'DeSoto'), (1287, 'LaMesa'), (1289, 'ElCajon'), ('...', '...'), (15081, 'LaCrosse'), (15301, 'Cincinnatiwilliam wolfe'), (15602, 'San Jose de las Lajas'), (15799, 'Puerto la Cruz'), (15908, 'Nuevo San Nicolas de los Garza'), (15910, 'San Nicolas de los Garza'), (15956, 'Hua-lien'), (15984, 'Neu-ulm'), (16395, 'San Juan de la Maguana'), (16401, 'Sabana de la Mar')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 4421 cities):

City Frequency

NULL 392

Chicago 361

Philadelphia 310

St. Louis 267

New York 236

Brooklyn 215

Los Angeles 214

Cincinnati 182

San Francisco 168

Baltimore 140

... ...

Yulee 1

Yuma 1

Zachary 1

Zaragoza 1

Zborov 1

Zealandia 1

Zebulon 1

Zeigler 1

Zeist 1

Zerbe 1

deathYear (year):

Year format: Error(s) found:

DQI #1 (Missing Data - Completeness):

8065 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (20, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (1872.0 : 2008.0)

deathMonth (month):

Month format: Error(s) found:

DQI #1 (Missing Data - Completeness):

8071 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (20, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution:

Month Frequency

January 755

December 742

April 731

October 723

March 719

November 712

May 699

September 674

February 663

August 651

July 635

June 632

deathDay (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

8071 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (20, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:31.0).

deathCountry (country):

Country format: Error(s) found:

DQI #1 (Missing Data - Completeness):

8105 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (20, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

4 Capitalization/Format issue(s) at index(es): [(2526, 'in a plane crash at sea'), (6743, 'Atlantic Ocean from N.Y. to Boston'), (8888, 'uSA'), (9771, 'Atlantic Ocean from N.Y. to P.R.')]

Frequency Distribution (showing top and bottom 10 of 27 countries):

Country Frequency

USA 8142

NULL 8105

CAN 58

Cuba 21

Mexico 16

Venezuela 15

D.R. 12

P.R. 10

Bahamas 4

France 4

... ...

England 1

Japan 1

Marianas Islands 1

North Korea (MIA) 1

Panama 1

Panama Canal Zone 1

Taiwan 1

V.I. 1

in a plane crash at sea 1

uSA 1

deathState (state):

State format: Error(s) found:

DQI #1 (Missing Data - Completeness):

8190 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (20, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

16 Extraneous data value(s) at index(es): [(186, '26'), (424, '30'), (667, '26'), (1081, '28'), (1096, '28'), (1722, '30'), (3410, '26'), (3651, '30'), (4680, '2'), (5185, '28'), (5199, '4'), (8332, '24'), (9748, '7'), (11991, '23'), (12312, '11'), (13221, '19')]

Frequency Distribution (showing top and bottom 10 of 67 states/provinces):

State Frequency

NULL 8190

CA 947

PA 736

NY 632

OH 536

FL 464

IL 429

MA 426

TX 423

MO 311

... ...

11 1

19 1

2 1

23 1

24 1

4 1

7 1

AK 1

BC 1

NB 1

deathCity (city):

City format: Error(s) found:

DQI #1 (Missing Data - Completeness):

8109 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (20, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

49 Capitalization/Format issue(s) at index(es): [(103, 'LaCrosse'), (451, 'LeRoy'), (1538, 'ElToro'), (2087, 'ElPaso'), (2590, 'ElPaso'), (2592, 'ElPaso'), (2700, 'LaMesa'), (2705, 'DeSoto'), (3202, 'LaGrange'), (3235, 'ElDorado'), ('...', '...'), (12574, 'ElMonte'), (12630, 'ElPaso'), (12851, 'DeSoto'), (13173, 'DuBois'), (13402, 'LaJolla'), (14033, 'LaJolla'), (14378, 'ElPaso'), (15075, 'LaJolla'), (15129, 'ElReno'), (15596, 'LaCrescenta')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 2344 cities):

City Frequency

NULL 8109

Philadelphia 224

Chicago 167

St.Louis 144

Los Angeles 122

New York 112

Cincinnati 104

Baltimore 102

Boston 88

Brooklyn 88

... ...

Yeadon 1

York Township 1

Yorktown 1

Yosemite 1

Yountville 1

Yuba City 1

Yucca Valley 1

Zachary 1

Zaleski 1

Zimmerman 1

nameFirst (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

48 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (48): [(653, 'NULL'), (1018, 'NULL'), (1239, 'NULL'), (1296, 'NULL'), (1616, 'NULL'), (1758, 'NULL'), (2639, 'NULL'), (3964, 'NULL'), (4004, 'NULL'), (4155, 'NULL'), ('...', '...'), (12713, 'NULL'), (12806, 'NULL'), (13030, 'NULL'), (13277, 'NULL'), (13530, 'NULL'), (13660, 'NULL'), (13924, 'NULL'), (15024, 'NULL'), (15141, 'NULL'), (15313, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

2 Capitalization/Format issue(s) at index(es): [(5936, 'LaTroy'), (15956, 'Chin-hui')]

Frequency Distribution:

nameFirst Frequency

Bill 544

John 451

Jim 431

Mike 397

Joe 381

Bob 337

George 299

Ed 280

Tom 276

Frank 267

... ...

Youngy 1

Yovanny 1

Yuniesky 1

Zachary 1

Zak 1

Zane 1

Zaza 1

Ziggy 1

Zinn 1

Zoilo 1

Range of Values: (A.J. to Zoilo)

nameLast (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

98 Capitalization/Format issue(s) at index(es): [(3341, 'de la Cruz'), (3342, 'de la Hoz'), (3343, 'de la Maza'), (3344, 'de la Rosa'), (3345, 'de la Rosa'), (3346, 'de la Rosa'), (3347, 'de los Santos'), (3348, 'de los Santos'), (3349, 'de los Santos'), (3365, 'DeArmond'), ('...', '...'), (15604, 'de los Santos'), (15770, 'DuBose'), (15851, 'DePaula'), (15925, 'DeJesus'), (15945, 'DePastino'), (15949, 'LaForest'), (16024, 'LaRoche'), (16032, 'DiNardo'), (16056, 'DeVore'), (16158, 'de la Rosa')] (displaying only the first and last 10 items)

Frequency Distribution:

nameLast Frequency

Smith 139

Johnson 100

Jones 90

Miller 80

Brown 76

Williams 73

Wilson 66

Davis 60

Moore 47

Taylor 45

... ...

Zuleta 1

Zupcic 1

Zupo 1

Zuvella 1

Zuverink 1

Zwilling 1

de la Cruz 1

de la Hoz 1

de la Maza 1

del Toro 1

Range of Values: (Aardsma to del Toro)

nameNote (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

15959 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (10, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

23 Extraneous data value(s) at index(es): [(315, 'born Peter William Jablonowski and legally changed name in 1933'), (703, 'Also Played One Game In 1918 Under Name Of Johnson'), (771, 'a.k.a. W.Thomas In 1871-1873'), (1305, 'a.k.a. Joseph Emley Josephs In 1875'), (1922, 'a.k.a. Frederick King In 1901'), (2560, 'a.k.a. Jack Mc Allister In 1902'), (4500, 'Also Played Under Name Of Wm. G. Morris 1 Game In 1915'), (6575, 'a.k.a. Al West For 1 Game In 1883'), (7306, 'Also Played Under Real Name Of Robert John Taggert In 1918'), (7634, 'a.k.a. Fred Brady In 1913'), ('...', '...'), (9889, 'a.k.a. Joseph Harley Bennett In 1918'), (10170, 'Played One Game In 1912 Under Name Of Costello'), (10478, 'a.k.a. Michael Joyce In 1901'), (11337, 'a.k.a. George Pepper Wilson In 1901'), (12243, 'Previously thought to be John Ryan (ryanjo02)'), (12667, 'Born with last name Sellman, Also Played Under Name Of Frank C. Williams 1871-75'), (13043, 'a.k.a. Robert M. Brown In 1914'), (14326, 'In 1984, played under real name of Jose Altagracia Gonzalez (Uribe)'), (15528, 'a.k.a. Zach Walton In 1918'), (16396, 'Played under the name Leo Nunez through 2011 season')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

410 Capitalization/Format issue(s) at index(es): [(9, 'born Harry Frederick Winbigler'), (18, 'Formerly thought to be Dave Abercrombie'), (21, 'also called Ted'), (99, 'born Edward Anshmedt'), (118, 'born Santo Anibal (Alcala)'), (135, 'born Manuel De Jesus (Alexander)'), (147, 'born Alvah Charles Elliott'), (151, 'Also played as Richie Allen'), (197, 'born Felipe Rojas (Alou)'), (198, 'born Jesus Maria Rojas (Alou)'), ('...', '...'), (15456, 'also Yellowhorse'), (15515, 'played as Ross Young'), (15526, 'born Albert Myron Zarski'), (15556, 'also Zeiser'), (15804, 'Record previously combined with Denny Driscoll'), (15805, 'Previously combined with Bill Morgan'), (15989, 'Formerly thought to be Mike Lehane'), (16183, 'also Jairo Garcia'), (16198, 'record previously given to other George Winkelman'), (16199, 'previously credited to Joe Yingling')] (displaying only the first and last 10 items)

Frequency Distribution:

nameNote Frequency

NULL 15959

also known as Joe 2

Also Emil Geis and Bill Geis 1

Also Known As Arthur Lee Whitehorn 1

Also Played One Game In 1918 Under Name Of Johnson 1

Also Played Under Name Of Albert May In 1872 1

Also Played Under Name Of Edward P. Bevens 1

Also Played Under Name Of Harry Zieber 1

Also Played Under Name Of Joseph Henry Greene 1

Also Played Under Name Of Ossee Schreck, born Schrecongost 1

... ...

previously spelled Claire 1

previously thought to be Edward L. 1

previously thought to be Kip Dowd 1

real last name Furrer 1

real last name Schultz 1

real name James Thomas Walsh 1

real name Pedro 1

real name is Eleutero Encarnacion 1

record previously given to other George Winkelman 1

went by Wayman 1

Range of Values: (Also Emil Geis and Bill Geis to went by Wayman)

nameGiven (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

741 Blank/Empty/Null/Missing value(s) at index(es): [(3, 'NULL'), (20, 'NULL'), (33, 'NULL'), (100, 'NULL'), (119, 'NULL'), (126, 'NULL'), (150, 'NULL'), (152, 'NULL'), (191, 'NULL'), (192, 'NULL'), ('...', '...'), (15474, 'NULL'), (15487, 'NULL'), (15511, 'NULL'), (15539, 'NULL'), (15559, 'NULL'), (15573, 'NULL'), (15575, 'NULL'), (15591, 'NULL'), (15804, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

14 Capitalization/Format issue(s) at index(es): [(2665, 'Jesus (De la Cruz)'), (5936, 'LaTroy'), (7987, 'DeWitt Wiley'), (8766, 'Albert DeGroot'), (8810, 'Manuel (DeJesus)'), (15607, 'Miguel (DePaula)'), (15614, 'Alexis DeJesus'), (15615, 'Francis LeBron'), (15678, 'Delvin DeWayne'), (15742, 'Desmond DeChone'), (15767, 'Eric DeBari'), (15895, 'Stephen LeCharles'), (15956, 'Chin-hui'), (15960, 'Rontrez DeMon')]

Frequency Distribution:

nameGiven Frequency

NULL 741

John Joseph 69

John 59

William Henry 50

William Joseph 41

William 38

James Edward 36

Michael Joseph 34

John Francis 32

John William 31

... ...

Zachary Quinn 1

Zachary Taylor 1

Zachary Thomas 1

Zakary Sebastien 1

Zane William 1

Zebulon Alexander 1

Zebulon Vance 1

Zerah Zequiel 1

Zinn Bertram 1

Zoilo Casanova (Rodriguez) 1

Range of Values: (A. Harry to Zoilo Casanova (Rodriguez))

nameNick (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

13632 Blank/Empty/Null/Missing value(s) at index(es): [(1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (7, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), ('...', '...'), (16396, 'NULL'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

3 Extraneous data value(s) at index(es): [(3327, '3-Dog'), (11991, 'Oh Henry!'), (15382, 'A-1')]

DQI #15 (Domain Violation - Accuracy):

10 Capitalization/Format issue(s) at index(es): [(462, 'BagPipes'), (4294, 'Death to Flying Things'), (5339, 'Baby-faced Assassin'), (6946, 'HoJo'), (7939, 'Long arm of the Law'), (8973, "Shake n' Bake"), (9661, 'Moonie; Moon-man'), (9788, 'Crossfire,Big Ebbie,Steam Engine in Boots'), (13242, "Slammin' Sammy, Say it Ain't Sosa"), (15058, "Sweet Swingin' Billy from Whistler")]

Frequency Distribution:

nameNick Frequency

NULL 13632

Lefty 125

Red 48

Dutch 32

Doc 26

Buck 17

Big Bill 15

Moose 15

Whitey 14

Big Ed 12

... ...

Zamboni 1

Zip 1

Zorro 1

the Animal 1

the Blade 1

the Bull 1

the Cat 1

the Chief 1

the Hacker 1

the Polish Prince 1

Range of Values: (3-Dog to the Polish Prince)

weight:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

786 Blank/Empty/Null/Missing value(s) at index(es): [(18, 'NULL'), (59, 'NULL'), (83, 'NULL'), (95, 'NULL'), (155, 'NULL'), (179, 'NULL'), (248, 'NULL'), (255, 'NULL'), (285, 'NULL'), (352, 'NULL'), ('...', '...'), (15550, 'NULL'), (15563, 'NULL'), (15567, 'NULL'), (15801, 'NULL'), (15803, 'NULL'), (15805, 'NULL'), (15988, 'NULL'), (15989, 'NULL'), (16198, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

Range of values: (65.0:320.0).

height:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

737 Blank/Empty/Null/Missing value(s) at index(es): [(18, 'NULL'), (59, 'NULL'), (83, 'NULL'), (95, 'NULL'), (159, 'NULL'), (179, 'NULL'), (248, 'NULL'), (255, 'NULL'), (285, 'NULL'), (328, 'NULL'), ('...', '...'), (15550, 'NULL'), (15563, 'NULL'), (15567, 'NULL'), (15801, 'NULL'), (15803, 'NULL'), (15805, 'NULL'), (15988, 'NULL'), (15989, 'NULL'), (16198, 'NULL'), (16200, 'NULL')] (displaying only the first and last 10 items)

Range of values: (43.0:83.0).

bats (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1193 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (1193): [(18, 'NULL'), (59, 'NULL'), (83, 'NULL'), (95, 'NULL'), (155, 'NULL'), (157, 'NULL'), (177, 'NULL'), (178, 'NULL'), (179, 'NULL'), (225, 'NULL'), ('...', '...'), (15578, 'NULL'), (15801, 'NULL'), (15803, 'NULL'), (15804, 'NULL'), (15805, 'NULL'), (15988, 'NULL'), (15989, 'NULL'), (16198, 'NULL'), (16199, 'NULL'), (16200, 'NULL')]

String range (lexicographical): (B : R)

Frequency Distribution:

bats Frequency

R 10039

L 4177

NULL 1193

B 998

throws (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

956 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (956): [(7, 'NULL'), (18, 'NULL'), (83, 'NULL'), (95, 'NULL'), (155, 'NULL'), (157, 'NULL'), (177, 'NULL'), (178, 'NULL'), (179, 'NULL'), (225, 'NULL'), ('...', '...'), (15578, 'NULL'), (15801, 'NULL'), (15803, 'NULL'), (15804, 'NULL'), (15805, 'NULL'), (15988, 'NULL'), (15989, 'NULL'), (16198, 'NULL'), (16199, 'NULL'), (16200, 'NULL')]

String range (lexicographical): (L : R)

Frequency Distribution:

throws Frequency

R 12377

L 3074

NULL 956

debut:

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (2): [(9095, 'NULL'), (16200, 'NULL')]

String range (lexicographical): (10/1/1874 : 9/9/2005)

Frequency Distribution:

debut Frequency

5/1/1884 36

5/1/1879 26

5/2/1882 26

4/17/1884 25

4/13/1954 24

4/19/1890 22

4/26/1995 21

4/15/1958 20

4/14/1942 19

4/17/1945 19

... ...

9/9/1954 1

9/9/1955 1

9/9/1956 1

9/9/1964 1

9/9/1970 1

9/9/1985 1

9/9/1991 1

9/9/1992 1

9/9/1997 1

9/9/2005 1

finalGame (final):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

788 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (788): [(30, 'NULL'), (56, 'NULL'), (76, 'NULL'), (126, 'NULL'), (159, 'NULL'), (178, 'NULL'), (253, 'NULL'), (264, 'NULL'), (301, 'NULL'), (363, 'NULL'), ('...', '...'), (16389, 'NULL'), (16390, 'NULL'), (16393, 'NULL'), (16396, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')]

String range (lexicographical): (10/1/1874 : 9/9/2006)

Frequency Distribution:

finalGame Frequency

NULL 788

9/28/2003 48

10/1/2000 47

10/3/1999 45

10/7/2001 43

10/5/1980 42

10/3/1993 41

9/29/1996 40

10/3/1915 39

10/6/1985 39

... ...

9/9/1978 1

9/9/1979 1

9/9/1980 1

9/9/1985 1

9/9/1986 1

9/9/1988 1

9/9/1989 1

9/9/1993 1

9/9/2000 1

9/9/2006 1

college:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

13350 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (13085): [(0, 'NULL'), (1, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (13, 'NULL'), (15, 'NULL'), ('...', '...'), (16396, 'NULL'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16406, 'NULL')]

Non-NULL values (265): [(15597, 'None'), (15602, 'None'), (15603, 'None'), (15604, 'None'), (15605, 'None'), (15606, 'None'), (15607, 'None'), (15608, 'None'), (15609, 'None'), (15610, 'None'), ('...', '...'), (16186, 'None'), (16187, 'None'), (16188, 'None'), (16189, 'None'), (16190, 'None'), (16191, 'None'), (16192, 'None'), (16193, 'None'), (16194, 'None'), (16195, 'None')]

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 882 unique value(s):

Category Frequency

NULL 13085

None 265

Arizona State 59

Stanford 46

Texas 45

UCLA 44

USC 41

Arizona 39

Florida State 39

Cal St. Fullerton 36

... ...

Wisconsin-Stevens Point 1

Wisconsin-Whitewater 1

YSU 1

Yakima Valley 1

Yale University 1

York 1

Youngstown State; Florida St. 1

ohio state 1

oklahoma 1

texas 1

lahman40ID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

810 Blank/Empty/Null/Missing value(s) at index(es): [(15597, 'NULL'), (15598, 'NULL'), (15599, 'NULL'), (15600, 'NULL'), (15601, 'NULL'), (15602, 'NULL'), (15603, 'NULL'), (15604, 'NULL'), (15605, 'NULL'), (15606, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

Alphanumeric range of values: (aaronha01 : zwilldu01)

lahman45ID (id):

ID column format: Error(s) found:

810 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (810): [(15597, 'NULL'), (15598, 'NULL'), (15599, 'NULL'), (15600, 'NULL'), (15601, 'NULL'), (15602, 'NULL'), (15603, 'NULL'), (15604, 'NULL'), (15605, 'NULL'), (15606, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')]

Alphanumeric range of values: (aaronha01 : zwilldu01)

retroID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (1): [(16200, 'NULL')]

Alphanumeric range of values: (aardd001 : zwild101)

holtzID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

802 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (802): 15597, 'NULL'), (15598, 'NULL'), (15599, 'NULL'), (15600, 'NULL'), (15601, 'NULL'), (15602, 'NULL'), (15603, 'NULL'), (15604, 'NULL'), (15605, 'NULL'), (15606, 'NULL'), ('...', '...'), (16397, 'NULL'), (16398, 'NULL'), (16399, 'NULL'), (16400, 'NULL'), (16401, 'NULL'), (16402, 'NULL'), (16403, 'NULL'), (16404, 'NULL'), (16405, 'NULL'), (16406, 'NULL')] (displaying only the first and last 10 items)

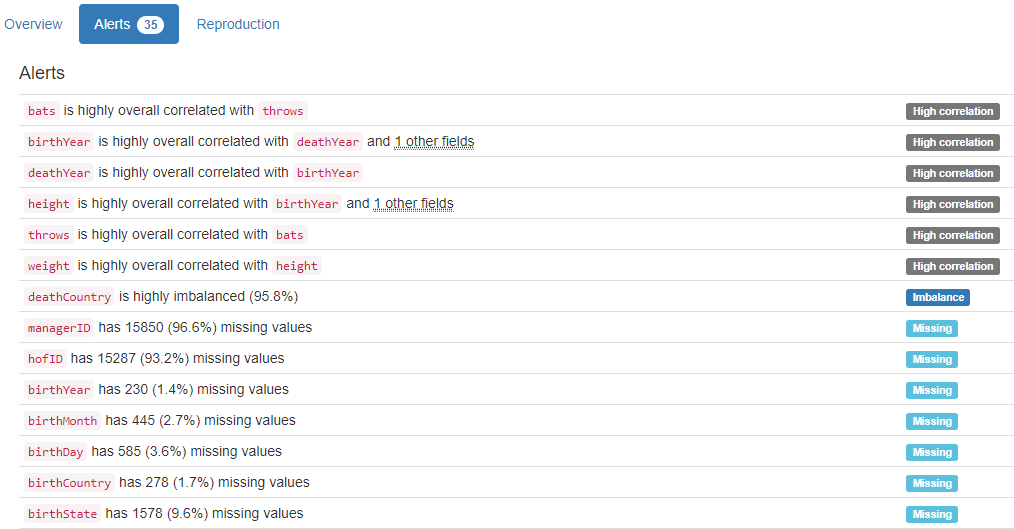
Alphanumeric range of values: (aaronha01 : zwilldu01)

bbrefID (id):

ID column format: All 16407 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (aardsda01 : zwilldu01)

YDATA Profiling results: It found correctly 25 of the 29 Missing values columns because they contain NULLs. However, it did not find all the Missing values for column 'college'. It just found correctly 13085 NULL cases of 13350 that were found by our system, because there are 265 values 'None'. Besides that, it did not find the Missing values on columns nameFirst (48), debut (2) and retroID (1). It also found that the PK column 'lahmanid' has unique values:



A screenshot of a computer

Description automatically generated

# DB.table: lahman\_2014.salaries – Sport – Type: Real

yearID|teamID|lgID|playerID (Composite Primary Key):

Composite Primary Key format: All 23111 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (1985|ATL|NL|barkele01 : 2012|WAS|NL|zimmery01)

salary:

Numerical >=0 format: All 23111 values are numerical and greater or equal to 0 in the range (0:33000000).

YDATA Profiling results: It just found out that the Composite PK with 4 columns, yearID|teamID|lgid|playerid has unique values:



# DB.table: lahman\_2014.teams – Sport – Type: Real

yearID|lgID|teamID (Composite Primary Key):

Composite Primary Key format: All 2715 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (1871|NA|BS1 : 2012|NL|WAS)

franchID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2668 Duplicate value(s) at index(es): [[0, 'BNA'], [1, 'CNA'], [2, 'CFC'], [4, 'NNA'], [5, 'PNA'], [7, 'TRO'], [8, 'OLY'], [9, 'BLC'], [11, 'BRA'], [12, 'BNA'], ('...', '...'), [2705, 'LAD'], [2706, 'FLA'], [2707, 'MIL'], [2708, 'NYM'], [2709, 'PHI'], [2710, 'PIT'], [2711, 'SDP'], [2712, 'SFG'], [2713, 'STL'], [2714, 'WSN']] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

2595 Uniqueness violation(s) at index(es): [(12, 'BNA'), (13, 'CFC'), (15, 'NNA'), (16, 'PNA'), (17, 'TRO'), (18, 'OLY'), (20, 'BLC'), (22, 'BRA'), (23, 'BNA'), (25, 'NNA'), ('...', '...'), (2705, 'LAD'), (2706, 'FLA'), (2707, 'MIL'), (2708, 'NYM'), (2709, 'PHI'), (2710, 'PIT'), (2711, 'SDP'), (2712, 'SFG'), (2713, 'STL'), (2714, 'WSN')] (displaying only the first and last 10 items)

Alphanumeric range of values: (ALT : WST)

divID (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1517 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (1507, 'NULL'), (1508, 'NULL'), (1509, 'NULL'), (1510, 'NULL'), (1511, 'NULL'), (1512, 'NULL'), (1513, 'NULL'), (1514, 'NULL'), (1515, 'NULL'), (1516, 'NULL')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):

1198 Duplicate value(s) at index(es): [[1517, 'E'], [1518, 'E'], [1519, 'W'], [1520, 'W'], [1521, 'E'], [1522, 'E'], [1523, 'W'], [1524, 'W'], [1525, 'E'], [1526, 'W'], ('...', '...'), [2705, 'W'], [2706, 'E'], [2707, 'C'], [2708, 'E'], [2709, 'E'], [2710, 'C'], [2711, 'W'], [2712, 'W'], [2713, 'C'], [2714, 'E']] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

1195 Uniqueness violation(s) at index(es): [(1518, 'E'), (1520, 'W'), (1521, 'E'), (1522, 'E'), (1523, 'W'), (1524, 'W'), (1525, 'E'), (1526, 'W'), (1527, 'W'), (1528, 'E'), ('...', '...'), (2705, 'W'), (2706, 'E'), (2707, 'C'), (2708, 'E'), (2709, 'E'), (2710, 'C'), (2711, 'W'), (2712, 'W'), (2713, 'C'), (2714, 'E')] (displaying only the first and last 10 items)

Alphanumeric range of values: (C : W)

Rank:

Numerical >=0 format: All 2715 values are numerical and greater or equal to 0 in the range (1:13).

G (integer):

Numerical format: All 2715 values are numerical in the range (6:165).

Ghome (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

549 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2675, 'NULL'), (2676, 'NULL'), (2677, 'NULL'), (2678, 'NULL'), (2679, 'NULL'), (2680, 'NULL'), (2681, 'NULL'), (2682, 'NULL'), (2683, 'NULL'), (2684, 'NULL')] (displaying only the first and last 10 items)

Range of values: (44.0:84.0).

W (integer):

Numerical format: All 2715 values are numerical in the range (0:116).

L (integer):

Numerical format: All 2715 values are numerical in the range (4:134).

DivWin (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1545 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2171, 'NULL'), (2172, 'NULL'), (2173, 'NULL'), (2174, 'NULL'), (2175, 'NULL'), (2176, 'NULL'), (2177, 'NULL'), (2178, 'NULL'), (2179, 'NULL'), (2180, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (N : Y)

WCWin (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2181 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2171, 'NULL'), (2172, 'NULL'), (2173, 'NULL'), (2174, 'NULL'), (2175, 'NULL'), (2176, 'NULL'), (2177, 'NULL'), (2178, 'NULL'), (2179, 'NULL'), (2180, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (N : Y)

LgWin (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

28 Blank/Empty/Null/Missing value(s) at index(es): [(2153, 'NULL'), (2154, 'NULL'), (2155, 'NULL'), (2156, 'NULL'), (2157, 'NULL'), (2158, 'NULL'), (2159, 'NULL'), (2160, 'NULL'), (2161, 'NULL'), (2162, 'NULL'), ('...', '...'), (2171, 'NULL'), (2172, 'NULL'), (2173, 'NULL'), (2174, 'NULL'), (2175, 'NULL'), (2176, 'NULL'), (2177, 'NULL'), (2178, 'NULL'), (2179, 'NULL'), (2180, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (N : Y)

WSWin (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

357 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2171, 'NULL'), (2172, 'NULL'), (2173, 'NULL'), (2174, 'NULL'), (2175, 'NULL'), (2176, 'NULL'), (2177, 'NULL'), (2178, 'NULL'), (2179, 'NULL'), (2180, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (N : Y)

R (integer):

Numerical format: All 2715 values are numerical in the range (24:1220).

AB (integer):

Numerical format: All 2715 values are numerical in the range (211:5781).

H (integer):

Numerical format: All 2715 values are numerical in the range (33:1783).

2B (integer):

Numerical format: All 2715 values are numerical in the range (3:376).

3B (integer):

Numerical format: All 2715 values are numerical in the range (0:150).

HR (integer):

Numerical format: All 2715 values are numerical in the range (0:264).

BB (integer):

Numerical format: All 2715 values are numerical in the range (0:835).

SO (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

120 Blank/Empty/Null/Missing value(s) at index(es): [(423, 'NULL'), (424, 'NULL'), (425, 'NULL'), (426, 'NULL'), (427, 'NULL'), (428, 'NULL'), (429, 'NULL'), (430, 'NULL'), (439, 'NULL'), (440, 'NULL'), ('...', '...'), (549, 'NULL'), (550, 'NULL'), (559, 'NULL'), (560, 'NULL'), (561, 'NULL'), (562, 'NULL'), (563, 'NULL'), (564, 'NULL'), (565, 'NULL'), (566, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:1529.0).

SB (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

144 Blank/Empty/Null/Missing value(s) at index(es): [(29, 'NULL'), (30, 'NULL'), (31, 'NULL'), (32, 'NULL'), (33, 'NULL'), (34, 'NULL'), (35, 'NULL'), (36, 'NULL'), (37, 'NULL'), (38, 'NULL'), ('...', '...'), (163, 'NULL'), (164, 'NULL'), (165, 'NULL'), (166, 'NULL'), (167, 'NULL'), (168, 'NULL'), (169, 'NULL'), (170, 'NULL'), (171, 'NULL'), (172, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:581.0).

CS (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

859 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (29, 'NULL'), ('...', '...'), (1181, 'NULL'), (1182, 'NULL'), (1191, 'NULL'), (1192, 'NULL'), (1193, 'NULL'), (1194, 'NULL'), (1195, 'NULL'), (1196, 'NULL'), (1197, 'NULL'), (1198, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:191.0).

HBP (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2325 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2315, 'NULL'), (2316, 'NULL'), (2317, 'NULL'), (2318, 'NULL'), (2319, 'NULL'), (2320, 'NULL'), (2321, 'NULL'), (2322, 'NULL'), (2323, 'NULL'), (2324, 'NULL')] (displaying only the first and last 10 items)

Range of values: (28.0:103.0).

SF (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2325 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2315, 'NULL'), (2316, 'NULL'), (2317, 'NULL'), (2318, 'NULL'), (2319, 'NULL'), (2320, 'NULL'), (2321, 'NULL'), (2322, 'NULL'), (2323, 'NULL'), (2324, 'NULL')] (displaying only the first and last 10 items)

Range of values: (24.0:75.0).

RA (integer):

Numerical format: All 2715 values are numerical in the range (34:1252).

ER (integer):

Numerical format: All 2715 values are numerical in the range (25:1023).

ERA (integer):

Numerical format: All 2715 values are numerical in the range (1:8).

CG (integer):

Numerical format: All 2715 values are numerical in the range (0:148).

SHO (integer):

Numerical format: All 2715 values are numerical in the range (0:32).

SV (integer):

Numerical format: All 2715 values are numerical in the range (0:68).

IPouts (integer):

Numerical format: All 2715 values are numerical in the range (162:4518).

HA (integer):

Numerical format: All 2715 values are numerical in the range (49:1993).

HRA (integer):

Numerical format: All 2715 values are numerical in the range (0:241).

BBA (integer):

Numerical format: All 2715 values are numerical in the range (0:827).

SOA (integer):

Numerical format: All 2715 values are numerical in the range (0:1404).

E (integer):

Numerical format: All 2715 values are numerical in the range (47:639).

DP (integer):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

317 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (365, 'NULL'), (366, 'NULL'), (367, 'NULL'), (368, 'NULL'), (369, 'NULL'), (370, 'NULL'), (371, 'NULL'), (372, 'NULL'), (373, 'NULL'), (374, 'NULL')] (displaying only the first and last 10 items)

Range of values: (18.0:217.0).

FP (integer):

Numerical format: All 2715 values are numerical in the range (0:0).

name:

Name format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

3 Extraneous data value(s) at index(es): [(1389, "Houston Colt .45's"), (1409, "Houston Colt .45's"), (1429, "Houston Colt .45's")]

Frequency Distribution:

name Frequency

Cincinnati Reds 122

Pittsburgh Pirates 122

Philadelphia Phillies 121

St. Louis Cardinals 113

Chicago White Sox 112

Detroit Tigers 112

Chicago Cubs 110

Boston Red Sox 105

New York Yankees 100

Cleveland Indians 98

... ...

Rockford Forest Citys 1

Seattle Pilots 1

St. Louis Perfectos 1

St. Louis Red Stockings 1

St. Paul White Caps 1

Toledo Blue Stockings 1

Toledo Maumees 1

Washington Blue Legs 1

Washington Statesmen 1

Wilmington Quicksteps 1

Range of Values: (Altoona Mountain City to Worcester Ruby Legs)

park:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

34 Blank/Empty/Null/Missing value(s) at index(es): [(145, 'NULL'), (146, 'NULL'), (147, 'NULL'), (148, 'NULL'), (149, 'NULL'), (150, 'NULL'), (151, 'NULL'), (152, 'NULL'), (153, 'NULL'), (154, 'NULL'), ('...', '...'), (604, 'NULL'), (605, 'NULL'), (606, 'NULL'), (623, 'NULL'), (624, 'NULL'), (625, 'NULL'), (627, 'NULL'), (628, 'NULL'), (629, 'NULL'), (630, 'NULL')]

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 213 unique value(s):

Category Frequency

Wrigley Field 99

Sportsman's Park IV 90

Comiskey Park 80

Fenway Park II 79

Forbes Field 60

Crosley Field 58

Polo Grounds IV 58

Shibe Park 58

Dodger Stadium 55

Yankee Stadium I 51

... ...

South Street Park 1

Speranza Park 1

Sportsman's Park IV/Busch Stadium II 1

Star Park II 1

Sun Life Stadium 1

Troy Ball Club Grounds 1

Union Base-Ball Grounds 1

Washington Park II 1

Waverly Fairgrounds 1

Wrigley Field (LA) 1

attendance:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

309 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (2645, 'NULL'), (2646, 'NULL'), (2647, 'NULL'), (2648, 'NULL'), (2649, 'NULL'), (2650, 'NULL'), (2651, 'NULL'), (2652, 'NULL'), (2653, 'NULL'), (2654, 'NULL')] (displaying only the first and last 10 items)

Range of values: (6088.0:4483350.0).

BPF (integer):

Numerical format: All 2715 values are numerical in the range (60:129).

PPF (integer):

Numerical format: All 2715 values are numerical in the range (60:141).

teamIDBR (team):

Categorical format: Error(s) found:

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 101 unique value(s):

Category Frequency

CHC 137

CIN 136

STL 134

PHI 133

PIT 131

BOS 120

CLE 117

CHW 112

DET 112

NYY 110

... ...

PHQ 1

RES 1

RIC 1

ROC 1

ROK 1

SEP 1

SLR 1

STP 1

WES 1

WIL 1

teamIDlahman45 (team):

Categorical format: Error(s) found:

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 148 unique value(s):

Category Frequency

CHN 137

PHI 130

PIT 126

CIN 123

SLN 121

BOS 112

CHA 112

CLE 112

DET 112

NYA 110

... ...

SR2 1

TL1 1

TL2 1

WIL 1

WS4 1

WS5 1

WS6 1

WS7 1

WS9 1

WSU 1

teamIDretro (team):

Categorical format: Error(s) found:

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 149 unique value(s):

Category Frequency

CHN 137

PHI 130

PIT 126

CIN 123

SLN 121

BOS 112

CHA 112

CLE 112

DET 112

NYA 110

... ...

SR2 1

TL1 1

TL2 1

WIL 1

WS4 1

WS5 1

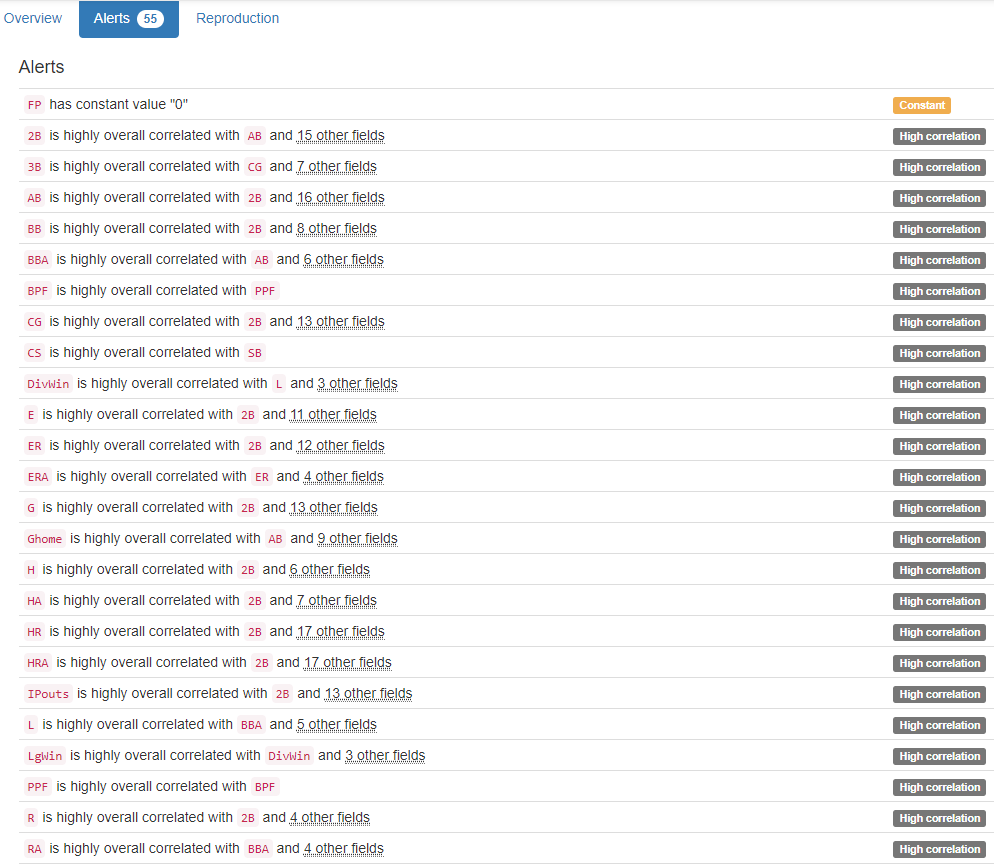
WS6 1

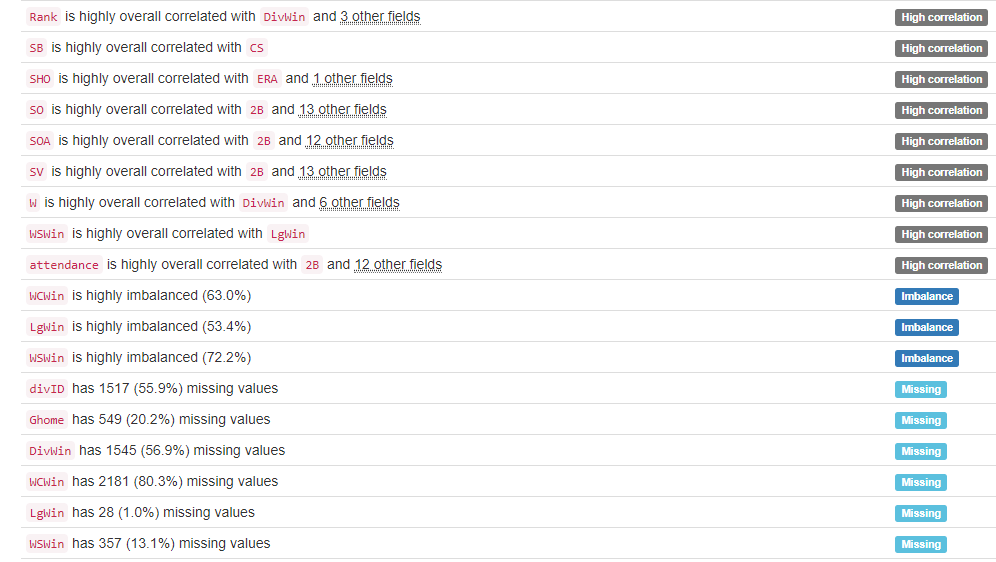
WS7 1

WS9 1

WSU 1

YDATA Profiling results: It found the same 14 Missing values columns because they contain NULLs. It also found that the column FP has constant value "0", and the Composite PK columns yearID|lgID|teamID has unique values:







# DB.table: legalActs.legalacts – Government – Type: Real

This database table’s content is in the Russian alphabet.

id:

Primary Key format: All 580049 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (4241 : 2332734)

hash:

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2 Duplicate value(s) at index(es): [[4759, '5e14f1a3d7cbbe58091a1bebb307fb3d'], [5095, '5e14f1a3d7cbbe58091a1bebb307fb3d']]

DQI #19 (Uniqueness Violation - Uniqueness):

1 Uniqueness violation(s) at index(es): [(5095, '5e14f1a3d7cbbe58091a1bebb307fb3d')]

Alphanumeric range of values: (000001fad84142961932d0e8fce7c5ca : fffff80bdf4bf46f26740a4cab71fb22)

update (timestamp):

Datetime format: All 580049 datetime values are valid in the YYYYMMDD format in the range 2012/07/13 14:03 to 2012/08/04 15:06.

Court:

Categorical format: Error(s) found:

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 177 unique value(s):

Category Frequency

Софийски градски съд 37477

Районен съд - Варна 30529

Софийски районен съд 21745

Районен съд - Пловдив 20342

Окръжен съд - Варна 18409

Районен съд - Бургас 16742

Районен съд - Русе 13672

Районен съд - Ст.Загора 13600

Районен съд - Плевен 12895

Районен съд - Благоевград 11654

... ...

Окръжен съд - Монтана 204

Военен съд - Сливен 79

Военен съд - Варна 62

Специализиран наказателен съд 57

Районен съд - Пирдоп 43

Военен съд - Плевен 28

Апелативен съд - Велико Търново 5

Военен съд - София 3

Върховен административен съд 2

Районен съд - Сандански 1

CaseKind (kind):

Categorical format: All 580049 values are correctly categorical.

Categorical format with 25 unique value(s):

Category Frequency

Гражданско дело 187992

НОХД 90266

АНД 87020

Въззивно гражданско дело 57266

Административно дело 31254

Въззивно частно гражданско дело 20595

Търговско дело 18238

ВНОХД 17868

ЧНД 17748

КАНД 11918

... ...

Брачно дело 2559

ВНЧХД 1840

ВАНД 1269

Фирмено дело 1158

Частно административно дело 900

КЧАНД 572

Частно търговско дело 485

Касационно административно дело 234

Касационно частно административно дело 89

Търговско дело по несъстоятелност 81

CaseNumber (number):

Numerical format: All 580049 values are numerical in the range (1:32767).

ActYear (year):

Year format: All 580049 values are numerical and valid in the range [1800, 2100].

Actual range of values: (1933 : 2012)

Judge:

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

474 Blank/Empty/Null/Missing value(s) at index(es): [(95673, 'NULL'), (96412, 'NULL'), (96416, 'NULL'), (96422, 'NULL'), (96427, 'NULL'), (96431, 'NULL'), (96434, 'NULL'), (96435, 'NULL'), (96439, 'NULL'), (96440, 'NULL'), ('...', '...'), (497608, 'NULL'), (497693, 'NULL'), (504927, 'NULL'), (509205, 'NULL'), (523352, 'NULL'), (560108, 'NULL'), (567412, 'NULL'), (567414, 'NULL'), (576372, 'NULL'), (577096, 'NULL')] (displaying only the first and last 10 items)

DQI #5 (Extraneous Data - Consistency, Uniqueness):

28516 Extraneous data value(s) at index(es): [(587, '1; Мариана Хитева, Румяна Цонева, Дочка Върбева'), (588, '4; Красимир Коларов, Емил Митев, Нестор Спасов'), (589, '9; Георги Чамбов, Красимир Коларов, Емил Митев'), (590, '3; Красимира Ангелова, Венета Махмудиева, Севдалин Божиков'), (591, '1; Мариана Хитева, Румяна Цонева, Дочка Върбева'), (592, '2; Евдокия Кемалова, Веселина Симеонова, Галя Геогиева'), (600, '3; Недялка Василева, Станко Кидиков, Христо Белев'), (601, '3; Недялка Василева, Станко Кидиков, Христо Белев'), (602, '3; Недялка Василева, Станко Кидиков, Христо Белев'), (603, '9; Георги Чамбов, Красимир Коларов, Емил Митев'), ('...', '...'), (578762, 'ГО 1-1 състав, АСЕН А. ВОДЕНИЧАРОВ'), (578763, 'ГО 1-1 състав, АСЕН А. ВОДЕНИЧАРОВ'), (578764, 'ГО 1-1 състав, АСЕН А. ВОДЕНИЧАРОВ'), (578766, 'НО 5 състав, ГЕОРГИ В. УШЕВ, ДАЯНА К. ЦВЕТАНОВА'), (578767, 'ГО 1-14 състав, КРАСИМИР Н. МАЗГАЛОВ'), (578768, 'ТО VI-3 състав, ЛИЛИЯ Ж. ИЛИЕВА'), (578771, 'ТО VI-13 състав, ДАНИЕЛА А. МАРЧЕВА'), (578772, 'ТО VI-8 състав, ИВАЙЛО С. МЛАДЕНОВ'), (578773, 'ТО VI-2 състав, АНЖЕЛИНА Д. ХРИСТОВА'), (578780, 'ГО 1-14 състав, КРАСИМИР Н. МАЗГАЛОВ')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

330681 Capitalization/Format issue(s) at index(es): [(0, 'V състав, ГЕОРГИ Д. ВАСИЛЕВ'), (1, 'III състав, ГАЛИНА И. ВЪЛЧАНОВА'), (2, 'II състав, НИКОЛАЙ Е. ИВАНОВ'), (3, 'II състав, НИКОЛАЙ Е. ИВАНОВ'), (4, 'VI състав, НЕВЕНА И. НЕСТОРОВА'), (5, 'VI състав, ВАЛЕНТИНА ПЕТРОВА'), (6, 'V състав, КРАСИМИРА Т. ДОНЕВА'), (7, 'III състав, ГАЛИНА И. ВЪЛЧАНОВА'), (8, 'VI състав, ВАЛЕНТИНА ПЕТРОВА'), (9, 'I състав, СВЕТЛА Р. ДИМИТРОВА'), ('...', '...'), (580026, 'III състав, ПЕТЯ ТОПАЛОВА'), (580027, 'II състав, КРАСИМИР СЕМОВ'), (580028, 'II състав, КАЛИН К. БАТАЛСКИ'), (580029, 'I състав, ИВЕТА П. ТАКОВА'), (580030, 'I състав, ИВЕТА П. ТАКОВА'), (580044, 'I състав, ТАТЯНА Г. БЕТОВА, МАРИАНА К. ТОДОРОВА, СТЕФАН А. ДАНЧЕВ'), (580045, 'I състав, РЕНИ М. СПАРТАНСКА, ЖАНЕТА Д. ГЕОРГИЕВА, ВЕСЕЛА Л. САХАТЧИЕВА'), (580046, 'I състав, РУМЕН П. ЛАЗАРОВ'), (580047, 'I състав, КАЛОЯН В. ГЕРГОВ, ГЕОРГИ К. ГРЪНЧАРОВ, СИЛВИЯ Ц. КРЪСТЕВА'), (580048, 'I състав, ЖАНЕТА Д. ГЕОРГИЕВА, МЕТОДИ Н. ЗДРАВКОВ, ЕКАТЕРИНА Т. ГЕОРГИЕВА')] (displaying only the first and last 10 items)

Frequency Distribution:

Judge Frequency

Мария Тодорова 2371

Серафимка Мадолева 2244

Росен Василев 1832

Василка Шаламанова 1721

Владимир Ковачев 1638

Иван Петков 1511

Саша Алексова 1265

V състав, ЗЛАТКА Т. ИЛИЕВА 1179

VII състав, ТОДОР Г. МИНОВ 1176

VI състав, ЗЛАТКО Д. МАЗНИКОВ 1171

... ...

смесен; Цветана Михова, Красимир Коларов, Нестор Спасов 1

смесен; Цветана Михова, Красимир Коларов, Симеон Захариев 1

смесен; Цветана Михова, Красимир Коларов, Стела Дандарова 1

смесен; Цветана Михова, Красимира Ангелова, Красимир Коларов 1

смесен; Цветана Михова, Мариана Хитева, Красимир Коларов 1

смесен; Цветана Михова, Недялка Василева, Мариана Хитева 1

смесен; Цветана Михова, Нестор Спасов, Симеон Захариев 1

смесен; Цветана Михова, Севдалин Божиков, Галина Арнаудова 1

смесен; Цветана Михова, Севдалин Божиков, Нестор Спасов 1

станчо димов 1

Range of Values: ( НАДЯ УЗУНОВА to цвета желязкова)

ActKind (kind):

Categorical format: All 580049 values are correctly categorical.

Categorical format with 7 unique value(s):

Category Frequency

Решение 365298

Определение 115400

Споразумение 42512

Присъда 39136

Заповед 6271

Протокол 6042

Разпореждане 5390

ActNumber (number):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

187900 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580034, 'NULL'), (580035, 'NULL'), (580036, 'NULL'), (580037, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580040, 'NULL'), (580041, 'NULL'), (580042, 'NULL'), (580043, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:32767.0).

StartDate (date):

Date format: All 580049 date values are valid in the YYYYMMDD format in the range 1999-07-09 to 2012-07-27.

LegalDate (date):

Date format: Error(s) found:

DQI #1 (Missing Data - Completeness):

293636 Blank/Empty/Null/Missing value(s) at index(es): [(12, 'NULL'), (13, 'NULL'), (15, 'NULL'), (18, 'NULL'), (19, 'NULL'), (20, 'NULL'), (21, 'NULL'), (26, 'NULL'), (27, 'NULL'), (29, 'NULL'), ('...', '...'), (580030, 'NULL'), (580031, 'NULL'), (580032, 'NULL'), (580035, 'NULL'), (580037, 'NULL'), (580044, 'NULL'), (580045, 'NULL'), (580046, 'NULL'), (580047, 'NULL'), (580048, 'NULL')] (displaying only the first and last 10 items)

Date range: 2006-07-03 to 2012-12-02

Status:

Categorical format: All 580049 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

Влязъл в сила 275395

Не е влязъл в сила 266867

Потвърден 16389

Отменен 13579

Изменен 7819

ActLink (tinyint):

Numerical format: All 580049 values are numerical in the range (0:1).

MotiveDate (date):

Date format: Error(s) found:

DQI #1 (Missing Data - Completeness):

530677 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580037, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580041, 'NULL'), (580042, 'NULL'), (580043, 'NULL'), (580044, 'NULL'), (580045, 'NULL'), (580047, 'NULL'), (580048, 'NULL')] (displaying only the first and last 10 items)

Date range: 1999-07-21 to 2012-07-27

MotiveLink (tinyint):

Numerical format: All 580049 values are numerical in the range (0:1).

HighCourt (court):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

465545 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580031, 'NULL'), (580033, 'NULL'), (580034, 'NULL'), (580035, 'NULL'), (580036, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580043, 'NULL'), (580047, 'NULL'), (580048, 'NULL')]

Categorical format with 60 unique value(s):

Category Frequency

NULL 465545

Върховен касационен съд 19023

Върховен административен съд 11057

Окръжен съд - Варна 7701

Апелативен съд - София 5789

Окръжен съд - Пловдив 4449

Административен съд - Варна 3984

Окръжен съд - Бургас 3408

Административен съд - Бургас 3039

Апелативен съд - Варна 2803

... ...

Окръжен съд - Търговище 540

Административен съд - Видин 504

Административен съд - Кърджали 474

Административен съд - Разград 472

Административен съд - Габрово 467

Административен съд - Търговище 465

Софийски окръжен съд 172

Административен съд - София-област 98

Военно-апелативен съд 56

Апелативен специализиран наказателен съд 11

OutNumber (number):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

465545 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580031, 'NULL'), (580033, 'NULL'), (580034, 'NULL'), (580035, 'NULL'), (580036, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580043, 'NULL'), (580047, 'NULL'), (580048, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:32767.0).

YearHigherCourt (year):

Year format: Error(s) found:

DQI #1 (Missing Data - Completeness):

465545 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580031, 'NULL'), (580033, 'NULL'), (580034, 'NULL'), (580035, 'NULL'), (580036, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580043, 'NULL'), (580047, 'NULL'), (580048, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (2005.0 : 2012.0)

TypeOfDocument (type):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

465545 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580031, 'NULL'), (580033, 'NULL'), (580034, 'NULL'), (580035, 'NULL'), (580036, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580043, 'NULL'), (580047, 'NULL'), (580048, 'NULL')]

Categorical format with 4 unique value(s):

Category Frequency

NULL 465545

Писмо 114404

Писмо - предложение за възобновяване 72

Писмо - молба за опр. срок при бавност 28

SendDate (date):

Date format: Error(s) found:

DQI #1 (Missing Data - Completeness):

465545 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580031, 'NULL'), (580033, 'NULL'), (580034, 'NULL'), (580035, 'NULL'), (580036, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580043, 'NULL'), (580047, 'NULL'), (580048, 'NULL')] (displaying only the first and last 10 items)

Date range: 2005-06-29 to 2012-07-27

ResultOfAppeal (result):

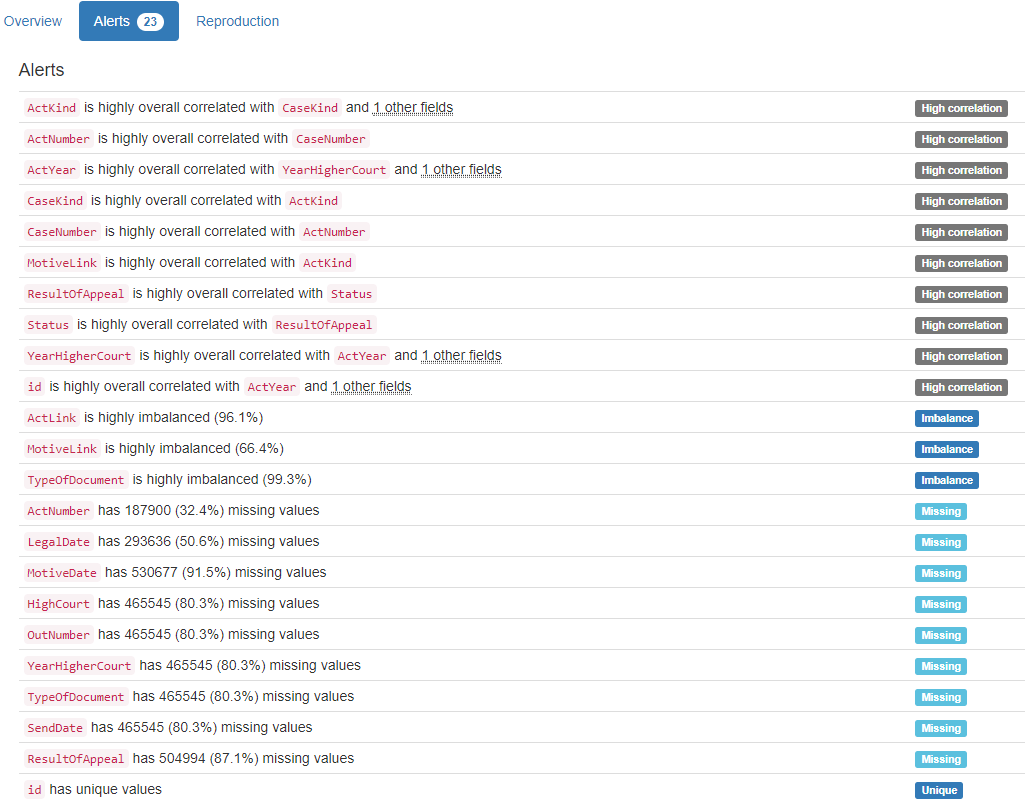
String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

504994 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (580035, 'NULL'), (580036, 'NULL'), (580037, 'NULL'), (580038, 'NULL'), (580039, 'NULL'), (580043, 'NULL'), (580044, 'NULL'), (580046, 'NULL'), (580047, 'NULL'), (580048, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (Изменен : Частично отменено и обезсилено, поради наличие на касационните основания по чл.209, т.3 АПК)

YDATA Profiling results: It found correctly 9 of the 10 Missing values columns because they contain NULLs. It did not find all the 474 Missing values for column 'judge' even though they are also NULLs. It also found that the PK column 'id' is Unique.



# DB.table: Mondial.city – Geography – Type: Real

Name|Province (Composite Primary Key):

Composite Primary Key format: All 3111 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (Aachen|Nordrhein Westfalen : s Hertogenbosch|Noord Brabant)

Country (id):

Foreign Key format: All 3111 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Alphanumeric range of values: (A : ZW)

References: country.Code

Population:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

465 Blank/Empty/Null/Missing value(s) at index(es): [(2, 'NULL'), (8, 'NULL'), (11, 'NULL'), (12, 'NULL'), (21, 'NULL'), (22, 'NULL'), (30, 'NULL'), (32, 'NULL'), (33, 'NULL'), (35, 'NULL'), ('...', '...'), (3048, 'NULL'), (3054, 'NULL'), (3063, 'NULL'), (3065, 'NULL'), (3066, 'NULL'), (3069, 'NULL'), (3080, 'NULL'), (3088, 'NULL'), (3106, 'NULL'), (3110, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:10229262.0).

Longitude:

Longitude format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2310 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (3101, 'NULL'), (3102, 'NULL'), (3103, 'NULL'), (3104, 'NULL'), (3105, 'NULL'), (3106, 'NULL'), (3107, 'NULL'), (3108, 'NULL'), (3109, 'NULL'), (3110, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (-175.2 : 178.0)

Latitude:

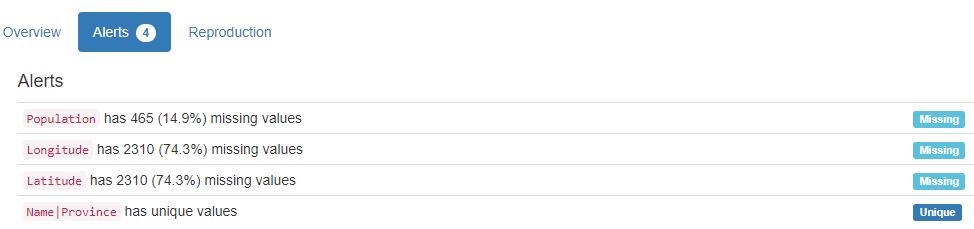
Latitude format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2310 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), ('...', '...'), (3101, 'NULL'), (3102, 'NULL'), (3103, 'NULL'), (3104, 'NULL'), (3105, 'NULL'), (3106, 'NULL'), (3107, 'NULL'), (3108, 'NULL'), (3109, 'NULL'), (3110, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (-54.5 : 78.13)

YDATA Profiling results: It found the same 3 Missing values columns because they contain NULLs. It also found that the Composite PK Name|Province is Unique.



# DB.table: Mondial.country – Geography – Type: Real

Name (country\_name):

Country format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

1 Capitalization/Format issue(s) at index(es): [(35, 'Cote dIvoire')]

Frequency Distribution (showing top and bottom 10 of 238 countries):

Country Frequency

Afghanistan 1

Albania 1

Algeria 1

American Samoa 1

Andorra 1

Angola 1

Anguilla 1

Antigua and Barbuda 1

Argentina 1

Armenia 1

... ...

Venezuela 1

Vietnam 1

Virgin Islands 1

Wallis and Futuna 1

West Bank 1

Western Sahara 1

Yemen 1

Zaire 1

Zambia 1

Zimbabwe 1

Code (id):

Primary Key format: All 238 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (A : ZW)

Capital:

City format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2 Blank/Empty/Null/Missing value(s) at index(es): [(69, 'NULL'), (226, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

6 Capitalization/Format issue(s) at index(es): [(5, 'Andorra la Vella'), (52, 'Dar es Salaam'), (118, 'Fort-de-France'), (167, 'Port-au-Prince'), (202, 'NDjamena'), (210, 'Port-of-Spain')]

Frequency Distribution (showing top and bottom 10 of 236 cities):

City Frequency

Kingston 2

NULL 2

Abu Dhabi 1

Abuja 1

Accra 1

Adamstown 1

Addis Ababa 1

Agana 1

Algier 1

Almaty 1

... ...

Washington 1

Wellington 1

West Island 1

Willemstad 1

Windhoek 1

Yamoussoukro 1

Yaounde 1

Yaren 1

Yerevan 1

Zagreb 1

Province:

State format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2 Blank/Empty/Null/Missing value(s) at index(es): [(69, 'NULL'), (226, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

17 Capitalization/Format issue(s) at index(es): [(2, 'Antigua and Barbuda'), (21, 'Bosnia and Herzegovina'), (30, 'Ciudad de la Habana'), (35, 'Cote dIvoire'), (85, 'Budapest (munic.)'), (101, 'Almaty (munic.)'), (104, 'Saint Kitts and Nevis'), (128, 'Maputo (munic.)'), (194, 'Saint Pierre and Miquelon'), (196, 'Sao Tome and Principe'), (197, 'al Khartum'), (201, 'Dushanbe (munic.)'), (204, 'Beijing (munic.)'), (210, 'Trinidad and Tobago'), (211, 'Turks and Caicos Islands'), (221, 'Wallis and Futuna'), (231, 'Saint Vincent and the Grenadines')]

Frequency Distribution (showing top and bottom 10 of 234 states/provinces):

State Frequency

Distrito Federal 4

NULL 2

Abu Dhabi 1

Afghanistan 1

Ahal 1

Albania 1

Algeria 1

Almaty (munic.) 1

American Samoa 1

Andorra 1

... ...

Vienna 1

Vietnam 1

Virgin Islands 1

Wallis and Futuna 1

Warszwaskie 1

Western Sahara 1

Yangon 1

Yemen 1

Zimbabwe 1

al Khartum 1

Area:

Numerical >=0 format: All 238 values are numerical and greater or equal to 0 in the range (0.44:17075200.0).

Population:

Numerical >=0 format: All 238 values are numerical and greater or equal to 0 in the range (48:1210004956).

YDATA Profiling results: It just found out that the PK column 'code' has unique values but did not find the Missing values on columns Capital (2), and province (2).

A white background with black lines

Description automatically generated

# DB.table: Mondial.politics – Geography – Type: Real

Country (id):

Primary Key format: All 238 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Alphanumeric range of values: (A : ZW)

Independence (end):

Date format: Error(s) found:

DQI #1 (Missing Data - Completeness):

55 Blank/Empty/Null/Missing value(s) at index(es): [(4, 'NULL'), (5, 'NULL'), (8, 'NULL'), (10, 'NULL'), (16, 'NULL'), (27, 'NULL'), (32, 'NULL'), (34, 'NULL'), (38, 'NULL'), (40, 'NULL'), ('...', '...'), (190, 'NULL'), (194, 'NULL'), (198, 'NULL'), (204, 'NULL'), (211, 'NULL'), (218, 'NULL'), (221, 'NULL'), (226, 'NULL'), (230, 'NULL'), (232, 'NULL')] (displaying only the first and last 10 items)

DQI #6 (Outdated Temporal Data - Timeliness):

11 Date value(s) not in [1800-2100] period at index(es): [(50, '1492-01-01'), (65, '1719-01-23'), (85, '1001-01-01'), (120, '1419-01-01'), (138, '1768-01-01'), (141, '1579-01-01'), (146, '1650-01-01'), (147, '1140-01-01'), (179, '0301-01-01'), (203, '1238-01-01'), (215, '1776-07-04')]

Date range: 0301-01-01 to 2008-02-17

Dependent (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

200 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (9, 'NULL'), (11, 'NULL'), (12, 'NULL'), ('...', '...'), (227, 'NULL'), (228, 'NULL'), (229, 'NULL'), (230, 'NULL'), (231, 'NULL'), (233, 'NULL'), (234, 'NULL'), (235, 'NULL'), (236, 'NULL'), (237, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Alphanumeric range of values: (AUS : USA)

References: country.Code

Government:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

3 Blank/Empty/Null/Missing value(s) at index(es): [(69, 'NULL'), (193, 'none'), (226, 'NULL')]

Categorical format with 74 unique value(s):

Category Frequency

republic 86

parliamentary democracy 20

constitutional monarchy 15

British Overseas Territories 9

federal republic 7

Communist state 5

emerging democracy 5

monarchy 4

overseas department of France 4

British crown dependency 3

... ...

republic within Commonwealth 1

republican parliamentary democracy 1

self governing parliamentary government in free association with New Zealand 1

self governing territory in free association with New Zealand 1

territory of Norway administered by the Ministry of Industry 1

theocratic republic 1

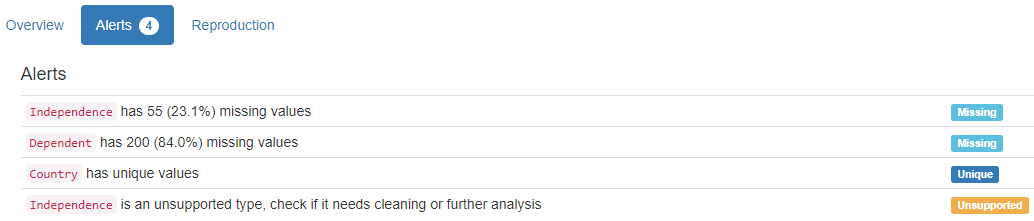
transitional government nominally a multiparty democracy with a strong presidential system 1

transitional previously ruling military junta 1

unincorporated and unorganized territory of the US 1

unitary republic 1

YDATA Profiling results: It found 2 of the 3 Missing values columns because they contain NULLs. However, it did not find all Missing values for column 'Government' (2 NULL and 1 'None'). It also found that the PK column 'Country' is Unique.



# DB.table: Mondial.population – Geography – Type: Real

Country (id):

Primary Key format: All 238 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Alphanumeric range of values: (A : ZW)

Population\_Growth (population):

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

9 Blank/Empty/Null/Missing value(s) at index(es): [(38, 'NULL'), (61, 'NULL'), (105, 'NULL'), (140, 'NULL'), (144, 'NULL'), (151, 'NULL'), (190, 'NULL'), (217, 'NULL'), (232, 'NULL')]

Range of values: (-3.3:16.49).

Infant\_Mortality (float):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

11 Blank/Empty/Null/Missing value(s) at index(es): [(38, 'NULL'), (61, 'NULL'), (105, 'NULL'), (125, 'NULL'), (140, 'NULL'), (144, 'NULL'), (151, 'NULL'), (190, 'NULL'), (198, 'NULL'), (217, 'NULL'), (232, 'NULL')]

Range of values: (2.2:149.7).

YDATA Profiling results: It found the same 2 Missing values columns because they contain NULLs. It also found that the PK column 'Country ' is Unique.

A screenshot of a computer

Description automatically generated

# DB.table: Mondial.province – Geography – Type: Real

Name|Country (Composite Primary Key):

Composite Primary Key format: All 1450 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (AG|CH : ash Sharqiyah|SUD)

Population:

Numerical >=0 format: All 1450 values are numerical and greater or equal to 0 in the range (0:206611600).

Area:

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

210 Blank/Empty/Null/Missing value(s) at index(es): [(9, 'NULL'), (18, 'NULL'), (31, 'NULL'), (34, 'NULL'), (37, 'NULL'), (49, 'NULL'), (58, 'NULL'), (60, 'NULL'), (65, 'NULL'), (87, 'NULL'), ('...', '...'), (1350, 'NULL'), (1363, 'NULL'), (1364, 'NULL'), (1370, 'NULL'), (1375, 'NULL'), (1390, 'NULL'), (1395, 'NULL'), (1406, 'NULL'), (1421, 'NULL'), (1446, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:3379680.0).

Capital:

City format: Error(s) found:

DQI #1 (Missing Data - Completeness):

9 Blank/Empty/Null/Missing value(s) at index(es): [(3, 'NULL'), (17, 'NULL'), (23, 'NULL'), (94, 'NULL'), (373, 'NULL'), (457, 'NULL'), (1062, 'NULL'), (1333, 'NULL'), (1395, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

22 Capitalization/Format issue(s) at index(es): [(2, 'LAquila'), (58, 'Andorra la Vella'), (93, 'ad Damir'), (264, 'NDjamena'), (266, 'Chalons en Champagne'), (339, 'Dar es Salaam'), (340, 'al Fasher'), (394, 'Shibin el Kom'), (516, 'Port-au-Prince'), (610, 'Kafr el Sheikh'), ('...', '...'), (711, 'Victoria de las Tunas'), (798, 'Fort-de-France'), (915, 's Hertogenbosch'), (941, 'Newcastle upon Tyne'), (1045, 'Dar es Salaam'), (1096, 'Rostov no Donu'), (1160, 'Usti nad Labem'), (1310, 'Port-of-Spain'), (1324, 'Newcastle upon Tyne'), (1448, 's Gravenhage')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 1406 cities):

City Frequency

NULL 9

Chandigarh 3

Kingston 3

La Paz 3

Merida 3

San Juan 3

Trujillo 3

Alexandria 2

Almaty 2

Barcelona 2

... ...

Ziguinchor 1

Zonguldak 1

Zug 1

Zurich 1

Zwolle 1

ad Damir 1

al Fasher 1

al Ubayyid 1

s Gravenhage 1

s Hertogenbosch 1

CapProv (varchar):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

9 Blank/Empty/Null/Missing value(s) at index(es): [(3, 'NULL'), (17, 'NULL'), (23, 'NULL'), (94, 'NULL'), (373, 'NULL'), (457, 'NULL'), (1062, 'NULL'), (1333, 'NULL'), (1395, 'NULL')]

String range (lexicographical): (AG : ash Sharqiyah)

YDATA Profiling results: It found only 1 of the 3 Missing values columns that contain NULLs. It did not find any Missing values for columns 'Capital' and CapProv (9 NULLs each). It also found that the Composite PK column 'Name|Country' is Unique.

A white background with black lines

Description automatically generated

# DB.table: northwind.Invoices – Retail – Type: Synth

ShipName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

837 Capitalization/Format issue(s) at index(es): [(4, 'Romero y tomillo'), (5, 'Romero y tomillo'), (6, 'Romero y tomillo'), (7, 'Du monde entier'), (8, 'Du monde entier'), (16, 'Bon app-'), (17, 'Bon app-'), (18, 'Bon app-'), (19, 'LILA-Supermercado'), (20, 'LILA-Supermercado'), ('...', '...'), (2137, 'Blondel pre et fils'), (2138, 'Blondel pre et fils'), (2139, 'Blondel pre et fils'), (2141, 'QUICK-Stop'), (2142, 'QUICK-Stop'), (2143, 'QUICK-Stop'), (2144, 'QUICK-Stop'), (2148, 'Rancho grande'), (2149, 'Rancho grande'), (2154, 'Blido Comidas preparadas')] (displaying only the first and last 10 items)

Frequency Distribution:

ShipName Frequency

Save-a-lot Markets 116

Ernst Handel 102

QUICK-Stop 86

Rattlesnake Canyon Grocery 71

Hungry Owl All-Night Grocers 55

Berglunds snabbkp 52

Frankenversand 48

Folk och f HB 45

HILARION-Abastos 45

Bon app- 44

... ...

Consolidated Holdings 7

The Big Cheese 7

Blido Comidas preparadas 6

France restauration 6

North/South 6

Spcialits du monde 6

GROSELLA-Restaurante 4

Alfreds Futterkiste 3

Centro comercial Moctezuma 2

Lazy K Kountry Store 2

Range of Values: (Alfred-s Futterkiste to Wolski Zajazd)

ShipAddress (address):

Street format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

339 Capitalization/Format issue(s) at index(es): [(19, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (20, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (21, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (22, 'ul. Filtrowa 68'), (23, 'ul. Filtrowa 68'), (39, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (40, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (41, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (57, 'Jardim das rosas n. 32'), (58, 'Jardim das rosas n. 32'), ('...', '...'), (2047, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (2079, 'Jardim das rosas n. 32'), (2087, '722 DaVinci Blvd.'), (2088, '722 DaVinci Blvd.'), (2089, '722 DaVinci Blvd.'), (2095, 'Carrera 22 con Ave. Carlos Soublette #8-35'), (2096, 'Carrera 22 con Ave. Carlos Soublette #8-35'), (2137, '24, place Klber'), (2138, '24, place Klber'), (2139, '24, place Klber')] (displaying only the first and last 10 items)

ShipCity (city):

City format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

70 Capitalization/Format issue(s) at index(es): [(42, 'rhus'), (43, 'rhus'), (44, 'rhus'), (45, 'rhus'), (46, 'rhus'), (206, 'rhus'), (207, 'rhus'), (208, 'rhus'), (283, 'rhus'), (284, 'rhus'), ('...', '...'), (1815, 'rhus'), (1863, 'Frankfurt a.M.'), (1902, 'Frankfurt a.M.'), (1903, 'Frankfurt a.M.'), (1904, 'Frankfurt a.M.'), (1911, 'rhus'), (1938, 'Frankfurt a.M.'), (1939, 'Frankfurt a.M.'), (2085, 'Frankfurt a.M.'), (2086, 'Frankfurt a.M.')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 70 cities):

City Frequency

Boise 116

Graz 102

Cunewalde 86

Rio de Janeiro 83

London 82

Sao Paulo 82

Mxico D.F. 72

Albuquerque 71

Cork 55

Lule 52

... ...

San Francisco 10

Torino 10

Elgin 9

Kirkland 9

Barcelona 8

Butte 8

Vancouver 8

Paris 6

Caracas 4

Walla Walla 2

ShipRegion (region):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1299 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (11, 'NULL'), ('...', '...'), (2141, 'NULL'), (2142, 'NULL'), (2143, 'NULL'), (2144, 'NULL'), (2145, 'NULL'), (2146, 'NULL'), (2147, 'NULL'), (2148, 'NULL'), (2149, 'NULL'), (2154, 'NULL')]

Categorical format with 20 unique value(s):

Category Frequency

NULL 1299

SP 120

ID 116

RJ 83

NM 71

Co. Cork 55

OR 52

WA 51

Tchira 45

BC 43

Nueva Esparta 35

Lara 34

Qubec 32

Essex 30

AK 24

Isle of Wight 23

WY 20

CA 10

MT 8

DF 4

ShipPostalCode (postal code):

Postal Code format: Error(s) found:

DQI #1 (Missing Data - Completeness):

55 Blank/Empty/Null/Missing value(s) at index(es): [(277, 'NULL'), (278, 'NULL'), (279, 'NULL'), (355, 'NULL'), (356, 'NULL'), (357, 'NULL'), (461, 'NULL'), (462, 'NULL'), (563, 'NULL'), (564, 'NULL'), ('...', '...'), (1808, 'NULL'), (2064, 'NULL'), (2065, 'NULL'), (2066, 'NULL'), (2067, 'NULL'), (2092, 'NULL'), (2093, 'NULL'), (2094, 'NULL'), (2097, 'NULL'), (2098, 'NULL')] (displaying only the first and last 10 items)

ShipCountry (country):

Country format: All 2155 country values are valid.

Frequency Distribution (showing top and bottom 10 of 21 countries):

Country Frequency

USA 352

Germany 328

Brazil 203

France 184

UK 135

Austria 125

Venezuela 118

Sweden 97

Canada 75

Mexico 72

... ...

Ireland 55

Finland 54

Spain 54

Italy 53

Switzerland 52

Denmark 46

Argentina 34

Portugal 30

Norway 16

Poland 16

CustomerID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2155 Duplicate value(s) at index(es): [[0, 'TORTU'], [1, 'TORTU'], [2, 'TORTU'], [3, 'TORTU'], [4, 'ROMEY'], [5, 'ROMEY'], [6, 'ROMEY'], [7, 'DUMON'], [8, 'DUMON'], [9, 'RATTC'], ('...', '...'), [2145, 'WARTH'], [2146, 'WARTH'], [2147, 'WARTH'], [2148, 'RANCH'], [2149, 'RANCH'], [2150, 'ISLAT'], [2151, 'ISLAT'], [2152, 'ISLAT'], [2153, 'ISLAT'], [2154, 'BOLID']] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

2066 Uniqueness violation(s) at index(es): [(1, 'TORTU'), (2, 'TORTU'), (3, 'TORTU'), (5, 'ROMEY'), (6, 'ROMEY'), (8, 'DUMON'), (10, 'RATTC'), (12, 'KOENE'), (13, 'KOENE'), (14, 'KOENE'), ('...', '...'), (2145, 'WARTH'), (2146, 'WARTH'), (2147, 'WARTH'), (2148, 'RANCH'), (2149, 'RANCH'), (2150, 'ISLAT'), (2151, 'ISLAT'), (2152, 'ISLAT'), (2153, 'ISLAT'), (2154, 'BOLID')] (displaying only the first and last 10 items)

Alphanumeric range of values: (ALFKI : WOLZA)

CustomerName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

787 Capitalization/Format issue(s) at index(es): [(4, 'Romero y tomillo'), (5, 'Romero y tomillo'), (6, 'Romero y tomillo'), (7, 'Du monde entier'), (8, 'Du monde entier'), (16, "Bon app'"), (17, "Bon app'"), (18, "Bon app'"), (19, 'LILA-Supermercado'), (20, 'LILA-Supermercado'), ('...', '...'), (2137, 'Blondesddsl pre et fils'), (2138, 'Blondesddsl pre et fils'), (2139, 'Blondesddsl pre et fils'), (2141, 'QUICK-Stop'), (2142, 'QUICK-Stop'), (2143, 'QUICK-Stop'), (2144, 'QUICK-Stop'), (2148, 'Rancho grande'), (2149, 'Rancho grande'), (2154, 'Blido Comidas preparadas')] (displaying only the first and last 10 items)

Frequency Distribution:

CustomerName Frequency

Save-a-lot Markets 116

Ernst Handel 102

QUICK-Stop 86

Rattlesnake Canyon Grocery 71

Hungry Owl All-Night Grocers 55

Berglunds snabbkp 52

Frankenversand 48

Folk och f HB 45

HILARION-Abastos 45

Bon app' 44

... ...

The Cracker Box 8

Consolidated Holdings 7

The Big Cheese 7

Blido Comidas preparadas 6

France restauration 6

North/South 6

Spcialits du monde 6

GROSELLA-Restaurante 4

Centro comercial Moctezuma 2

Lazy K Kountry Store 2

Range of Values: (Alfreds Futterkiste to Wolski Zajazd)

Address:

Street format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

339 Capitalization/Format issue(s) at index(es): [(19, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (20, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (21, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (22, 'ul. Filtrowa 68'), (23, 'ul. Filtrowa 68'), (39, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (40, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (41, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (57, 'Jardim das rosas n. 32'), (58, 'Jardim das rosas n. 32'), ('...', '...'), (2047, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (2079, 'Jardim das rosas n. 32'), (2087, '722 DaVinci Blvd.'), (2088, '722 DaVinci Blvd.'), (2089, '722 DaVinci Blvd.'), (2095, 'Carrera 22 con Ave. Carlos Soublette #8-35'), (2096, 'Carrera 22 con Ave. Carlos Soublette #8-35'), (2137, '24, place Klber'), (2138, '24, place Klber'), (2139, '24, place Klber')] (displaying only the first and last 10 items)

City:

City format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

70 Capitalization/Format issue(s) at index(es): [(42, 'rhus'), (43, 'rhus'), (44, 'rhus'), (45, 'rhus'), (46, 'rhus'), (206, 'rhus'), (207, 'rhus'), (208, 'rhus'), (283, 'rhus'), (284, 'rhus'), ('...', '...'), (1815, 'rhus'), (1863, 'Frankfurt a.M.'), (1902, 'Frankfurt a.M.'), (1903, 'Frankfurt a.M.'), (1904, 'Frankfurt a.M.'), (1911, 'rhus'), (1938, 'Frankfurt a.M.'), (1939, 'Frankfurt a.M.'), (2085, 'Frankfurt a.M.'), (2086, 'Frankfurt a.M.')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 69 cities):

City Frequency

Boise 116

London 112

Graz 102

Cunewalde 86

Rio de Janeiro 83

Sao Paulo 82

Mxico D.F. 72

Albuquerque 71

Cork 55

Lule 52

... ...

San Francisco 10

Torino 10

Elgin 9

Kirkland 9

Barcelona 8

Butte 8

Vancouver 8

Paris 6

Caracas 4

Walla Walla 2

Region:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1329 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (11, 'NULL'), ('...', '...'), (2141, 'NULL'), (2142, 'NULL'), (2143, 'NULL'), (2144, 'NULL'), (2145, 'NULL'), (2146, 'NULL'), (2147, 'NULL'), (2148, 'NULL'), (2149, 'NULL'), (2154, 'NULL')]

Categorical format with 19 unique value(s):

Category Frequency

NULL 1329

SP 120

ID 116

RJ 83

NM 71

Co. Cork 55

OR 52

WA 51

Tchira 45

BC 43

Nueva Esparta 35

Lara 34

Qubec 32

AK 24

Isle of Wight 23

WY 20

CA 10

MT 8

DF 4

PostalCode (postal code):

Postal Code format: Error(s) found:

DQI #1 (Missing Data - Completeness):

55 Blank/Empty/Null/Missing value(s) at index(es): [(277, 'NULL'), (278, 'NULL'), (279, 'NULL'), (355, 'NULL'), (356, 'NULL'), (357, 'NULL'), (461, 'NULL'), (462, 'NULL'), (563, 'NULL'), (564, 'NULL'), ('...', '...'), (1808, 'NULL'), (2064, 'NULL'), (2065, 'NULL'), (2066, 'NULL'), (2067, 'NULL'), (2092, 'NULL'), (2093, 'NULL'), (2094, 'NULL'), (2097, 'NULL'), (2098, 'NULL')] (displaying only the first and last 10 items)

Country:

Country format: All 2155 country values are valid.

Frequency Distribution (showing top and bottom 10 of 21 countries):

Country Frequency

USA 352

Germany 328

Brazil 203

France 184

UK 135

Austria 125

Venezuela 118

Sweden 97

Canada 75

Mexico 72

... ...

Ireland 55

Finland 54

Spain 54

Italy 53

Switzerland 52

Denmark 46

Argentina 34

Portugal 30

Norway 16

Poland 16

Salesperson (double):

Numerical format: All 2155 values are numerical in the range (0.0:0.0).

OrderID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2018 Duplicate value(s) at index(es): [[0, 10293], [1, 10293], [2, 10293], [3, 10293], [4, 10306], [5, 10306], [6, 10306], [7, 10311], [8, 10311], [9, 10316], ('...', '...'), [2144, 10745], [2145, 10750], [2146, 10750], [2147, 10750], [2148, 10828], [2149, 10828], [2150, 10829], [2151, 10829], [2152, 10829], [2153, 10829]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

1325 Uniqueness violation(s) at index(es): [(1, 10293), (2, 10293), (3, 10293), (5, 10306), (6, 10306), (8, 10311), (10, 10316), (12, 10325), (13, 10325), (14, 10325), ('...', '...'), (2139, 10566), (2142, 10745), (2143, 10745), (2144, 10745), (2146, 10750), (2147, 10750), (2149, 10828), (2151, 10829), (2152, 10829), (2153, 10829)] (displaying only the first and last 10 items)

Numeric range of values: (10248 : 11077)

OrderDate (datetime):

Datetime format: All 2155 datetime values are valid in the YYYYMMDD format in the range 1996/07/04 00:00 to 1998/05/06 00:00.

RequiredDate (datetime):

Datetime format: All 2155 datetime values are valid in the YYYYMMDD format in the range 1996/07/24 00:00 to 1998/06/11 00:00.

ShippedDate (datetime):

Datetime format: Error(s) found:

DQI #1 (Missing Data - Completeness):

73 Blank/Empty/Null/Missing value(s) at index(es): [(217, 'NULL'), (218, 'NULL'), (219, 'NULL'), (220, 'NULL'), (223, 'NULL'), (224, 'NULL'), (225, 'NULL'), (226, 'NULL'), (227, 'NULL'), (228, 'NULL'), ('...', '...'), (1976, 'NULL'), (1977, 'NULL'), (1978, 'NULL'), (2044, 'NULL'), (2045, 'NULL'), (2046, 'NULL'), (2047, 'NULL'), (2080, 'NULL'), (2081, 'NULL'), (2082, 'NULL')] (displaying only the first and last 10 items)

ShipperName (name):

Name format: All 2155 ShipperName values are valid.

Frequency Distribution:

ShipperName Frequency

United Package 864

Speedy Express 646

Federal Shipping 645

Range of Values: (Federal Shipping to United Package)

ProductID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2155 Duplicate value(s) at index(es): [[0, 18], [1, 24], [2, 63], [3, 75], [4, 30], [5, 53], [6, 54], [7, 42], [8, 69], [9, 41], ('...', '...'), [2145, 14], [2146, 45], [2147, 59], [2148, 20], [2149, 38], [2150, 2], [2151, 8], [2152, 13], [2153, 60], [2154, 52]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

2078 Uniqueness violation(s) at index(es): [(16, 18), (17, 41), (22, 31), (27, 14), (29, 26), (30, 31), (31, 13), (32, 62), (35, 72), (40, 30), ('...', '...'), (2145, 14), (2146, 45), (2147, 59), (2148, 20), (2149, 38), (2150, 2), (2151, 8), (2152, 13), (2153, 60), (2154, 52)] (displaying only the first and last 10 items)

Numeric range of values: (1 : 77)

ProductName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

377 Capitalization/Format issue(s) at index(es): [(2, 'Vegie-spread'), (10, 'Tarte au sucre'), (15, 'Mozzarella di Giovanni'), (25, 'Chartreuse verte'), (28, 'NuNuCa Nu-Nougat-Creme'), (32, 'Tarte au sucre'), (35, 'Mozzarella di Giovanni'), (41, 'Pt chinois'), (45, 'Rogede sild'), (46, 'Valkoinen suklaa'), ('...', '...'), (2067, 'Original Frankfurter grne Soe'), (2074, 'Zaanse koeken'), (2082, "Sirop d'rable"), (2083, 'NuNuCa Nu-Nougat-Creme'), (2085, 'Wimmers gute Semmelkndel'), (2087, 'Chartreuse verte'), (2089, 'Original Frankfurter grne Soe'), (2132, 'Vegie-spread'), (2144, 'Mozzarella di Giovanni'), (2146, 'Rogede sild')] (displaying only the first and last 10 items)

Frequency Distribution:

ProductName Frequency

Raclette Courdavault 54

Camembert Pierrot 51

Gorgonzola Telino 51

Guaran Fantstica 51

Gnocchi di nonna Alice 50

Tarte au sucre 48

Jack's New England Clam Chowder 47

Rhnbru Klosterbier 46

Chang 44

Pavlova 43

... ...

Grandma's Boysenberry Spread 12

Chef Anton's Gumbo Mix 10

Laughing Lumberjack Lager 10

Valkoinen suklaa 10

Schoggi Schokolade 9

Louisiana Hot Spiced Okra 8

Chocolade 6

Genen Shouyu 6

Gravad lax 6

Mishi Kobe Niku 5

Range of Values: (Alice Mutton to Zaanse koeken)

UnitPrice (price):

Numerical format: All 2155 values are numerical in the range (2.0:263.5).

Quantity:

Numerical format: All 2155 values are numerical in the range (1:130).

Discount (double):

Numerical format: All 2155 values are numerical in the range (0.0:0.0).

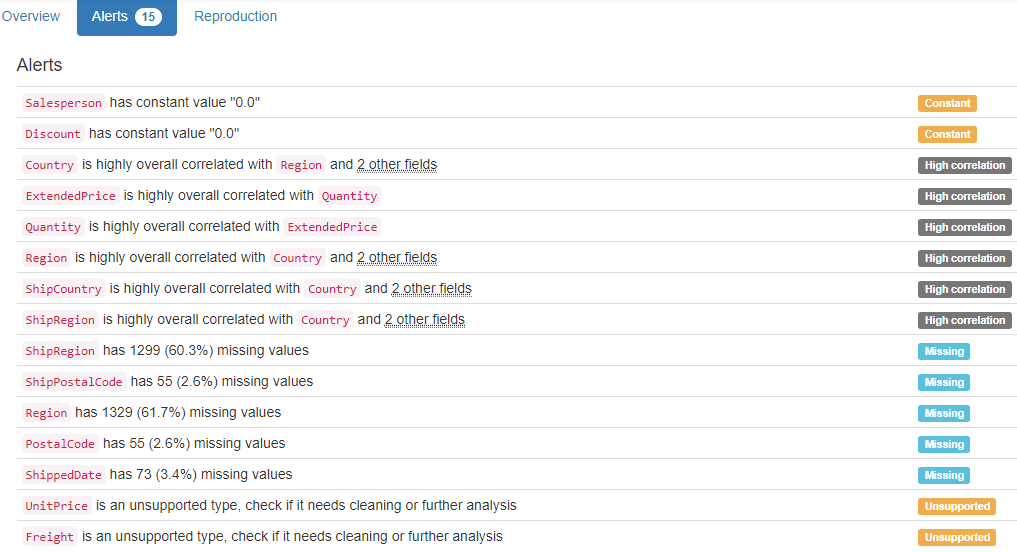
ExtendedPrice (price):

Numerical format: All 2155 values are numerical in the range (4.8:15810.0).

Freight (decimal):

Numerical format: All 2155 values are numerical in the range (0.02:1007.64).

YDATA Profiling results: It found the same 5 Missing values columns because they contain NULLs. It also found that the columns 'Salesperson' and 'Discount' have constant value "0", and it did not find any Unique values because there is no PK in this table:



# DB.table: northwind.Order Details Extended – Retail – Type: Synth – This is a Query, not a Table.

OrderID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2018 Duplicate value(s) at index(es): [[0, 10265], [2, 10294], [3, 10302], [4, 10319], [5, 10338], [6, 10339], [7, 10346], [8, 10415], [9, 10430], [10, 10431], ('...', '...'), [2144, 10685], [2146, 10710], [2147, 10714], [2148, 10755], [2149, 10762], [2150, 10837], [2151, 10875], [2152, 10928], [2153, 10956], [2154, 10977]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

1325 Uniqueness violation(s) at index(es): [(56, 10431), (69, 10607), (93, 10294), (116, 10658), (117, 10684), (119, 10728), (120, 10742), (126, 10895), (137, 11059), (140, 11077), ('...', '...'), (2144, 10685), (2146, 10710), (2147, 10714), (2148, 10755), (2149, 10762), (2150, 10837), (2151, 10875), (2152, 10928), (2153, 10956), (2154, 10977)] (displaying only the first and last 10 items)

Numeric range of values: (10248 : 11077)

ProductID (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

2155 Duplicate value(s) at index(es): [[0, 17], [1, 17], [2, 17], [3, 17], [4, 17], [5, 17], [6, 17], [7, 17], [8, 17], [9, 17], ('...', '...'), [2145, 47], [2146, 47], [2147, 47], [2148, 47], [2149, 47], [2150, 47], [2151, 47], [2152, 47], [2153, 47], [2154, 47]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

2078 Uniqueness violation(s) at index(es): [(1, 17), (2, 17), (3, 17), (4, 17), (5, 17), (6, 17), (7, 17), (8, 17), (9, 17), (10, 17), ('...', '...'), (2145, 47), (2146, 47), (2147, 47), (2148, 47), (2149, 47), (2150, 47), (2151, 47), (2152, 47), (2153, 47), (2154, 47)] (displaying only the first and last 10 items)

Numeric range of values: (1 : 77)

ProductName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

377 Capitalization/Format issue(s) at index(es): [(250, 'Chartreuse verte'), (251, 'Chartreuse verte'), (252, 'Chartreuse verte'), (253, 'Chartreuse verte'), (254, 'Chartreuse verte'), (255, 'Chartreuse verte'), (256, 'Chartreuse verte'), (257, 'Chartreuse verte'), (258, 'Chartreuse verte'), (259, 'Chartreuse verte'), ('...', '...'), (2145, 'Zaanse koeken'), (2146, 'Zaanse koeken'), (2147, 'Zaanse koeken'), (2148, 'Zaanse koeken'), (2149, 'Zaanse koeken'), (2150, 'Zaanse koeken'), (2151, 'Zaanse koeken'), (2152, 'Zaanse koeken'), (2153, 'Zaanse koeken'), (2154, 'Zaanse koeken')] (displaying only the first and last 10 items)

Frequency Distribution:

ProductName Frequency

Raclette Courdavault 54

Camembert Pierrot 51

Gorgonzola Telino 51

Guaran Fantstica 51

Gnocchi di nonna Alice 50

Tarte au sucre 48

Jack's New England Clam Chowder 47

Rhnbru Klosterbier 46

Chang 44

Pavlova 43

... ...

Grandma's Boysenberry Spread 12

Chef Anton's Gumbo Mix 10

Laughing Lumberjack Lager 10

Valkoinen suklaa 10

Schoggi Schokolade 9

Louisiana Hot Spiced Okra 8

Chocolade 6

Genen Shouyu 6

Gravad lax 6

Mishi Kobe Niku 5

Range of Values: (Alice Mutton to Zaanse koeken)

UnitPrice (price):

Numerical format: All 2155 values are numerical in the range (2.0:263.5).

Quantity:

Numerical format: All 2155 values are numerical in the range (1:130).

Discount (double):

Numerical format: All 2155 values are numerical in the range (0.0:0.0).

ExtendedPrice (price):

Numerical format: All 2155 values are numerical in the range (4.8:15810.0).

YDATA Profiling results: It just found that the columns 'Discount' has constant value "0", and it did not find any Unique values because there is no PK in this view:

A screenshot of a computer

Description automatically generated

# DB.table: northwind.Orders – Retail – Type: Synth

OrderID (id):

Primary Key format: All 830 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (10248 : 11077)

CustomerID (id):

Foreign Key format: All 830 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Alphanumeric range of values: (ALFKI : WOLZA)

References: Customers.CustomerID

EmployeeID (id):

Foreign Key format: All 830 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 9)

References: Employees.EmployeeID

OrderDate (datetime):

Datetime format: All 830 datetime values are valid in the YYYYMMDD format in the range 1996/07/04 00:00 to 1998/05/06 00:00.

RequiredDate (datetime):

Datetime format: All 830 datetime values are valid in the YYYYMMDD format in the range 1996/07/24 00:00 to 1998/06/11 00:00.

ShippedDate (datetime):

Datetime format: Error(s) found:

DQI #1 (Missing Data - Completeness):

21 Blank/Empty/Null/Missing value(s) at index(es): [(760, 'NULL'), (771, 'NULL'), (791, 'NULL'), (792, 'NULL'), (797, 'NULL'), (803, 'NULL'), (806, 'NULL'), (810, 'NULL'), (811, 'NULL'), (813, 'NULL'), ('...', '...'), (817, 'NULL'), (820, 'NULL'), (822, 'NULL'), (823, 'NULL'), (824, 'NULL'), (825, 'NULL'), (826, 'NULL'), (827, 'NULL'), (828, 'NULL'), (829, 'NULL')] (displaying only the first and last 10 items)

ShipVia (id):

Foreign Key format: All 830 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 3)

References: Shippers.ShipperID

Freight (decimal):

Numerical format: All 830 values are numerical in the range (0.02:1007.64).

ShipName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

319 Capitalization/Format issue(s) at index(es): [(0, 'Vins et alcools Chevalier'), (3, 'Victuailles en stock'), (4, 'Suprmes dlices'), (6, 'Chop-suey Chinese'), (9, 'HILARION-Abastos'), (11, 'Centro comercial Moctezuma'), (16, 'Folk och f HB'), (17, 'Blondel pre et fils'), (20, 'GROSELLA-Restaurante'), (25, 'QUICK-Stop'), ('...', '...'), (805, 'Piccolo und mehr'), (806, 'Cactus Comidas para llevar'), (807, 'HILARION-Abastos'), (812, 'Franchi S.p.A.'), (816, 'Save-a-lot Markets'), (817, 'LILA-Supermercado'), (823, 'LILA-Supermercado'), (825, 'Pericles Comidas clsicas'), (826, 'Simons bistro'), (828, 'Bon app-')] (displaying only the first and last 10 items)

Frequency Distribution:

ShipName Frequency

Save-a-lot Markets 31

Ernst Handel 30

QUICK-Stop 28

Folk och f HB 19

Hungry Owl All-Night Grocers 19

Berglunds snabbkp 18

HILARION-Abastos 18

Rattlesnake Canyon Grocery 18

Bon app- 17

Frankenversand 15

... ...

Consolidated Holdings 3

France restauration 3

Laughing Bacchus Wine Cellars 3

North/South 3

The Cracker Box 3

Trail-s Head Gourmet Provisioners 3

GROSELLA-Restaurante 2

Lazy K Kountry Store 2

Alfreds Futterkiste 1

Centro comercial Moctezuma 1

Range of Values: (Alfred-s Futterkiste to Wolski Zajazd)

ShipAddress (address):

Street format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

137 Capitalization/Format issue(s) at index(es): [(0, '59 rue de l-Abbaye'), (9, 'Carrera 22 con Ave. Carlos Soublette #8-35'), (16, 'kergatan 24'), (17, '24, place Klber'), (26, '59 rue de l-Abbaye'), (27, 'Via Ludovico il Moro 22'), (35, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (42, 'Av. dos Lusadas, 23'), (47, '59 rue de l-Abbaye'), (48, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), ('...', '...'), (753, 'kergatan 24'), (759, 'Estrada da sade n. 58'), (767, 'Erling Skakkes gate 78'), (794, 'Av. dos Lusadas, 23'), (796, 'ul. Filtrowa 68'), (802, 'kergatan 24'), (807, 'Carrera 22 con Ave. Carlos Soublette #8-35'), (817, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo'), (820, 'Alameda dos Canrios, 891'), (823, 'Carrera 52 con Ave. Bolvar #65-98 Llano Largo')] (displaying only the first and last 10 items)

ShipCity (city):

City format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

26 Capitalization/Format issue(s) at index(es): [(31, 'Frankfurt a.M.'), (36, 'Frankfurt a.M.'), (95, 'Frankfurt a.M.'), (119, 'rhus'), (151, 'rhus'), (217, 'rhus'), (249, 'Frankfurt a.M.'), (274, 'Frankfurt a.M.'), (286, 'Frankfurt a.M.'), (288, 'Frankfurt a.M.'), ('...', '...'), (496, 'rhus'), (521, 'rhus'), (524, 'Frankfurt a.M.'), (614, 'Frankfurt a.M.'), (643, 'Frankfurt a.M.'), (673, 'rhus'), (686, 'Frankfurt a.M.'), (698, 'rhus'), (746, 'rhus'), (822, 'Frankfurt a.M.')] (displaying only the first and last 10 items)

Frequency Distribution (showing top and bottom 10 of 70 cities):

City Frequency

Rio de Janeiro 34

London 33

Boise 31

Sao Paulo 31

Graz 30

Cunewalde 28

Mxico D.F. 28

Brcke 19

Cork 19

Albuquerque 18

... ...

Lille 5

Reims 5

Paris 4

San Francisco 4

Versailles 4

Butte 3

Kirkland 3

Vancouver 3

Caracas 2

Walla Walla 2

ShipRegion (region):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

507 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (7, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (15, 'NULL'), ('...', '...'), (812, 'NULL'), (814, 'NULL'), (819, 'NULL'), (821, 'NULL'), (822, 'NULL'), (824, 'NULL'), (825, 'NULL'), (826, 'NULL'), (827, 'NULL'), (828, 'NULL')]

Categorical format with 20 unique value(s):

Category Frequency

NULL 507

SP 49

RJ 34

ID 31

OR 28

Co. Cork 19

WA 19

NM 18

Tchira 18

BC 17

Lara 14

Essex 13

Qubec 13

Nueva Esparta 12

AK 10

Isle of Wight 10

WY 9

CA 4

MT 3

DF 2

ShipPostalCode (postal code):

Postal Code format: Error(s) found:

DQI #1 (Missing Data - Completeness):

19 Blank/Empty/Null/Missing value(s) at index(es): [(50, 'NULL'), (61, 'NULL'), (87, 'NULL'), (125, 'NULL'), (132, 'NULL'), (181, 'NULL'), (255, 'NULL'), (268, 'NULL'), (319, 'NULL'), (398, 'NULL'), (413, 'NULL'), (439, 'NULL'), (453, 'NULL'), (464, 'NULL'), (488, 'NULL'), (649, 'NULL'), (664, 'NULL'), (737, 'NULL'), (815, 'NULL')]

ShipCountry (country):

Country format: All 830 country values are valid.

Frequency Distribution (showing top and bottom 10 of 21 countries):

Country Frequency

Germany 122

USA 122

Brazil 83

France 77

UK 56

Venezuela 46

Austria 40

Sweden 37

Canada 30

Italy 28

... ...

Spain 23

Finland 22

Belgium 19

Ireland 19

Denmark 18

Switzerland 18

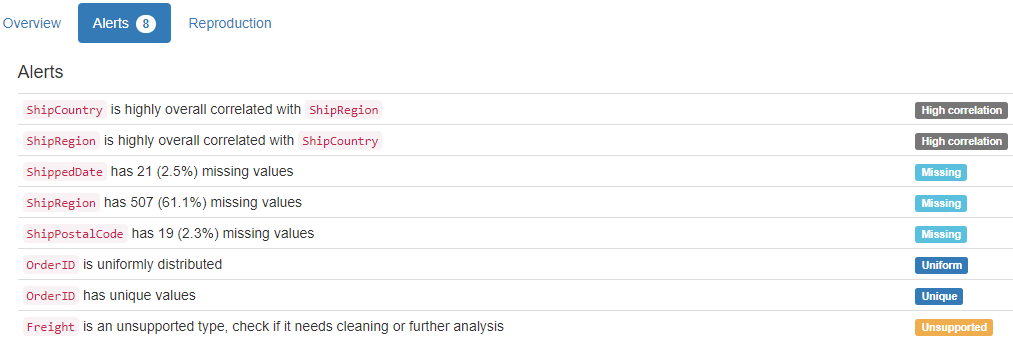
Argentina 16

Portugal 13

Poland 7

Norway 6

YDATA Profiling results: It found the same 3 Missing values columns because they contain NULLs. It also found that the PK column 'OrderID' is unique and uniformly distributed.



# DB.table: northwind.Products – Retail – Type: Synth

ProductID (id):

Primary Key format: All 77 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 77)

ProductName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

14 Capitalization/Format issue(s) at index(es): [(24, 'NuNuCa Nu-Nougat-Creme'), (36, 'Gravad lax'), (38, 'Chartreuse verte'), (44, 'Rogede sild'), (46, 'Zaanse koeken'), (49, 'Valkoinen suklaa'), (54, 'Pt chinois'), (55, 'Gnocchi di nonna Alice'), (60, "Sirop d'rable"), (61, 'Tarte au sucre'), (62, 'Vegie-spread'), (63, 'Wimmers gute Semmelkndel'), (71, 'Mozzarella di Giovanni'), (76, 'Original Frankfurter grne Soe')]

Frequency Distribution:

ProductName Frequency

Alice Mutton 1

Aniseed Syrup 1

Boston Crab Meat 1

Camembert Pierrot 1

Carnarvon Tigers 1

Chai 1

Chang 1

Chartreuse verte 1

Chef Anton's Cajun Seasoning 1

Chef Anton's Gumbo Mix 1

... ...

Teatime Chocolate Biscuits 1

Thringer Rostbratwurst 1

Tofu 1

Tourtire 1

Tunnbrd 1

Uncle Bob's Organic Dried Pears 1

Valkoinen suklaa 1

Vegie-spread 1

Wimmers gute Semmelkndel 1

Zaanse koeken 1

Range of Values: (Alice Mutton to Zaanse koeken)

SupplierID (id):

Foreign Key format: All 77 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 29)

References: Suppliers.SupplierID

CategoryID (id):

Foreign Key format: All 77 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 8)

References: Categories.CategoryID

QuantityPerUnit (quantity):

Numerical format: Error(s) found:

DQI #17 (Wrong Data Type - Consistency):

77 Non-numeric value(s) at index(es): [(0, '10 boxes x 20 bags'), (1, '24 - 12 oz bottles'), (2, '12 - 550 ml bottles'), (3, '48 - 6 oz jars'), (4, '36 boxes'), (5, '12 - 8 oz jars'), (6, '12 - 1 lb pkgs.'), (7, '12 - 12 oz jars'), (8, '18 - 500 g pkgs.'), (9, '12 - 200 ml jars'), ('...', '...'), (67, '10 boxes x 8 pieces'), (68, '10 kg pkg.'), (69, '24 - 355 ml bottles'), (70, '10 - 500 g pkgs.'), (71, '24 - 200 g pkgs.'), (72, '24 - 150 g jars'), (73, '5 kg pkg.'), (74, '24 - 0.5 l bottles'), (75, '500 ml'), (76, '12 boxes')] (displaying only the first and last 10 items)

No valid numeric values found.

UnitPrice (price):

Numerical format: All 77 values are numerical in the range (2.5:263.5).

UnitsInStock (stock):

Numerical format: All 77 values are numerical in the range (0:125).

UnitsOnOrder (units):

Numerical format: All 77 values are numerical in the range (0:100).

ReorderLevel (level):

Categorical format: All 77 values are correctly categorical.

Categorical format with 7 unique value(s):

Category Frequency

0 24

25 12

15 10

5 8

20 8

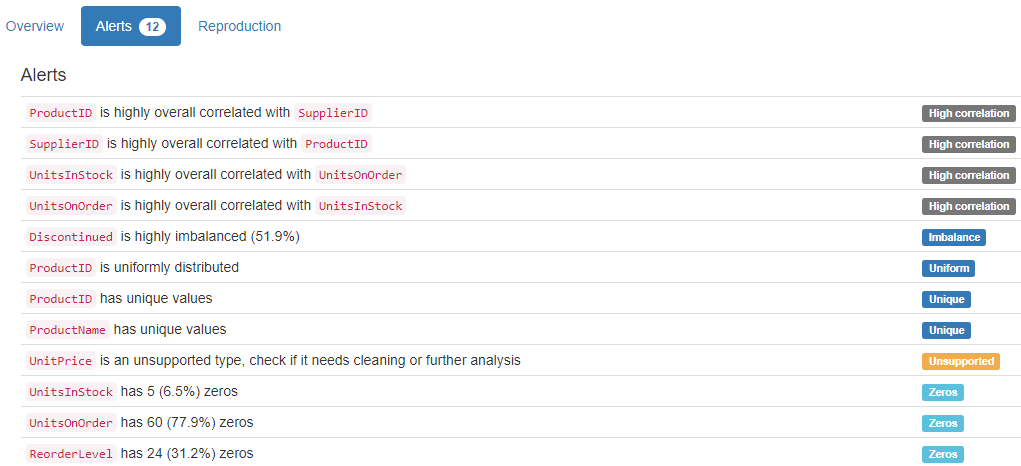
30 8

10 7

Discontinued (tinyint):

Numerical format: All 77 values are numerical in the range (0:1).

YDATA Profiling results: It just found out that the PK column ' ProductID' has unique values and is uniformly distributed:



# DB.table: northwind.Suppliers – Retail – Type: Synth

SupplierID (id):

Primary Key format: All 29 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 29)

CompanyName (name):

Name format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

8 Capitalization/Format issue(s) at index(es): [(10, 'Heli Swaren GmbH & Co. KG'), (12, 'Nord-Ost-Fisch Handelsgesellschaft mbH'), (13, 'Formaggi Fortini s.r.l.'), (17, 'Aux joyeux ecclsiastiques'), (23, "G'day, Mate"), (25, 'Pasta Buttini s.r.l.'), (27, 'Gai pturage'), (28, "Forts d'rables")]

Frequency Distribution:

CompanyName Frequency

Aux joyeux ecclsiastiques 1

Bigfoot Breweries 1

Cooperativa de Quesos 'Las Cabras' 1

Escargots Nouveaux 1

Exotic Liquids 1

Formaggi Fortini s.r.l. 1

Forts d'rables 1

G'day, Mate 1

Gai pturage 1

Grandma Kelly's Homestead 1

... ...

Norske Meierier 1

PB Knckebrd AB 1

Pasta Buttini s.r.l. 1

Pavlova, Ltd. 1

Plutzer Lebensmittelgromrkte AG 1

Refrescos Americanas LTDA 1

Specialty Biscuits, Ltd. 1

Svensk Sjfda AB 1

Tokyo Traders 1

Zaanse Snoepfabriek 1

Range of Values: (Aux joyeux ecclsiastiques to Zaanse Snoepfabriek)

ContactName (name):

Name format: All 29 ContactName values are valid.

Frequency Distribution:

ContactName Frequency

Anne Heikkonen 1

Antonio del Valle Saavedra 1

Beate Vileid 1

Carlos Diaz 1

Chandra Leka 1

Chantal Goulet 1

Charlotte Cooper 1

Cheryl Saylor 1

Dirk Luchte 1

Eliane Noz 1

... ...

Michael Bjrn 1

Niels Petersen 1

Peter Wilson 1

Petra Winkler 1

Regina Murphy 1

Robb Merchant 1

Shelley Burke 1

Sven Petersen 1

Wendy Mackenzie 1

Yoshi Nagase 1

Range of Values: (Anne Heikkonen to Yoshi Nagase)

ContactTitle (title):

Categorical format: All 29 values are correctly categorical.

Categorical format with 15 unique value(s):

Category Frequency

Sales Representative 6

Marketing Manager 5

Sales Manager 4

Accounting Manager 2

Order Administrator 2

Coordinator Foreign Markets 1

Export Administrator 1

International Marketing Mgr. 1

Marketing Representative 1

Owner 1

Product Manager 1

Purchasing Manager 1

Regional Account Rep. 1

Sales Agent 1

Wholesale Account Agent 1

Address:

Street format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

9 Extraneous street data value(s) at index(es): [(3, '9-8 Sekimai\r\nMusashino-shi'), (5, '92 Setsuko\r\nChuo-ku'), (6, '74 Rose St.\r\nMoonie Ponds'), (15, '3400 - 8th Avenue\r\nSuite 210'), (18, 'Order Processing Dept.\r\n2100 Paul Revere Blvd.'), (20, 'Lyngbysild\r\nFiskebakken 10'), (21, 'Verkoop\r\nRijnweg 22'), (23, "170 Prince Edward Parade\r\nHunter's Hill"), (27, 'Bat. B\r\n3, rue des Alpes')]

DQI #15 (Domain Violation - Accuracy):

2 Capitalization/Format issue(s) at index(es): [(12, 'Frahmredder 112a'), (25, 'Via dei Gelsomini, 153')]

City:

City format: All 29 city values are valid.

Frequency Distribution (showing top and bottom 10 of 29 cities):

City Frequency

Ann Arbor 1

Annecy 1

Bend 1

Berlin 1

Boston 1

Cuxhaven 1

Frankfurt 1

Gteborg 1

Lappeenranta 1

London 1

... ...

Ravenna 1

Salerno 1

Sandvika 1

Singapore 1

So Paulo 1

Ste-Hyacinthe 1

Stockholm 1

Sydney 1

Tokyo 1

Zaandam 1

Region:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

20 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (3, 'NULL'), (5, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (12, 'NULL'), (13, 'NULL'), (14, 'NULL'), (16, 'NULL'), (17, 'NULL'), (19, 'NULL'), (20, 'NULL'), (21, 'NULL'), (22, 'NULL'), (25, 'NULL'), (26, 'NULL'), (27, 'NULL')]

Categorical format with 9 unique value(s):

Category Frequency

NULL 20

Qubec 2

Asturias 1

LA 1

MA 1

MI 1

NSW 1

OR 1

Victoria 1

PostalCode (postal code):

Postal Code format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

2 Short length alphanumeric value(s) at index(es): [(3, '100'), (5, '545')]

Country:

Country format: All 29 country values are valid.

Frequency Distribution:

Country Frequency

USA 4

France 3

Germany 3

Australia 2

Canada 2

Italy 2

Japan 2

Sweden 2

UK 2

Brazil 1

Denmark 1

Finland 1

Netherlands 1

Norway 1

Singapore 1

Spain 1

Phone:

Phone format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

3 Incorrect telephone number format issue(s) at index(es): [(17, '(1) 03.83.00.68'), (26, '85.57.00.07'), (27, '38.76.98.06')]

Fax:

Phone format: Error(s) found:

DQI #1 (Missing Data - Completeness):

16 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (7, 'NULL'), (9, 'NULL'), (10, 'NULL'), (11, 'NULL'), (14, 'NULL'), (15, 'NULL'), (16, 'NULL'), (19, 'NULL'), (22, 'NULL'), (24, 'NULL'), (26, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

2 Incorrect telephone number format issue(s) at index(es): [(17, '(1) 03.83.00.62'), (27, '38.76.98.58')]

HomePage (home page):

URL format: Error(s) found:

DQI #1 (Missing Data - Completeness):

24 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), (12, 'NULL'), ('...', '...'), (18, 'NULL'), (19, 'NULL'), (20, 'NULL'), (21, 'NULL'), (22, 'NULL'), (24, 'NULL'), (25, 'NULL'), (26, 'NULL'), (27, 'NULL'), (28, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

5 Invalid URL format issue(s) at index(es): [(1, '#CAJUN.HTM#'), (5, "Mayumi's (on the World Wide Web)#[http://www.microsoft.com/accessdev/sampleapps/mayumi.htm#](http://www.microsoft.com/accessdev/sampleapps/mayumi.htm)"), (11, 'Plutzer (on the World Wide Web)#[http://www.microsoft.com/accessdev/sampleapps/plutzer.htm#](http://www.microsoft.com/accessdev/sampleapps/plutzer.htm)'), (13, '#FORMAGGI.HTM#'), (23, "G'day Mate (on the World Wide Web)#[http://www.microsoft.com/accessdev/sampleapps/gdaymate.htm#](http://www.microsoft.com/accessdev/sampleapps/gdaymate.htm)")]

YDATA Profiling results: It found the same 3 Missing values columns because they contain NULLs. It also found that the PK column 'SupplierID' is unique and is uniformly distributed.

A white page with black lines

Description automatically generated

# DB.table: PTE.pte\_atm – Medicine – Type: Real

drug\_id (id):

Foreign Key format: Error(s) found:

DQI #17 (Wrong Data Type - Consistency):

2593 Inconsistent length in alphanumeric value(s) at index(es): [(307, 'd10'), (308, 'd10'), (309, 'd10'), (310, 'd10'), (311, 'd10'), (312, 'd10'), (313, 'd10'), (314, 'd10'), (315, 'd10'), (316, 'd10'), ('...', '...'), (9179, 'd9'), (9180, 'd9'), (9181, 'd9'), (9182, 'd9'), (9183, 'd9'), (9184, 'd9'), (9185, 'd9'), (9186, 'd9'), (9187, 'd9'), (9188, 'd9')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Alphanumeric range of values: (d1 : d99)

References: pte\_drug.drug\_id

atom\_id (id):

Primary Key format: All 9189 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (d100\_1 : d9\_9)

atom\_type (type):

Categorical format: All 9189 values are correctly categorical.

Categorical format with 24 unique value(s):

Category Frequency

h 4101

c 3471

o 734

n 394

cl 281

s 96

br 45

p 24

na 16

f 7

... ...

as 1

ba 1

ca 1

hg 1

k 1

mn 1

pb 1

se 1

te 1

zn 1

Arg3 (varchar):

String format: All 9189 string values are valid.

String range (lexicographical): (1 : 96)

Arg4 (decimal):

Numerical format: All 9189 values are numerical in the range (-0.812:2.007).

YDATA Profiling results: It just found out that the PK column 'atom\_id' has unique values:

A screenshot of a computer

Description automatically generated

# DB.table: PTE.pte\_drug – Medicine – Type: Real

drug\_id (id):

Primary Key format: All 340 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (d1 : d99)

YDATA Profiling results: It just found out that the PK column 'drug\_id' has unique values:



# DB.table: sakila.address – Retail – Type: Synth

address\_id (id):

Primary Key format: All 603 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 605)

address:

Street format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

15 Extraneous street data value(s) at index(es): [(24, '262 A Corua (La Corua) Parkway'), (36, '127 Purnea (Purnia) Manor'), (39, '334 Munger (Monghyr) Lane'), (57, '1964 Allappuzha (Alleppey) Street'), (75, '89 Allappuzha (Alleppey) Manor'), (202, '1149 A Corua (La Corua) Boulevard'), (222, '1244 Allappuzha (Alleppey) Place'), (225, '810 Palghat (Palakkad) Boulevard'), (240, '647 A Corua (La Corua) Street'), (325, '1427 A Corua (La Corua) Place'), (473, '434 Ourense (Orense) Manor'), (518, '568 Dhule (Dhulia) Loop'), (521, '680 A Corua (La Corua) Manor'), (526, '600 Purnea (Purnia) Avenue'), (563, '1229 Varanasi (Benares) Manor')]

DQI #15 (Domain Violation - Accuracy):

14 Capitalization/Format issue(s) at index(es): [(0, '47 MySakila Drive'), (1, '28 MySQL Boulevard'), (185, '533 al-Ayn Boulevard'), (198, '1792 Valle de la Pascua Place'), (209, '1304 s-Hertogenbosch Way'), (283, '1006 Santa Brbara dOeste Manor'), (340, '124 al-Manama Way'), (364, '114 Jalib al-Shuyukh Manor'), (378, '1519 Santiago de los Caballeros Loop'), (423, '1866 al-Qatif Avenue'), (494, '775 ostka Drive'), (509, '1152 al-Qatif Lane'), (525, '1176 Southend-on-Sea Manor'), (552, '1764 Jalib al-Shuyukh Parkway')]

address2 (address):

Street format: Error(s) found:

603 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'Empty'), (5, 'Empty'), (6, 'Empty'), (7, 'Empty'), (8, 'Empty'), (9, 'Empty'), ('...', '...'), (593, 'Empty'), (594, 'Empty'), (595, 'Empty'), (596, 'Empty'), (597, 'Empty'), (598, 'Empty'), (599, 'Empty'), (600, 'Empty'), (601, 'Empty'), (602, 'Empty')] (displaying only the first and last 10 items)

Frequency Distribution:

Street Frequency

Empty 599

NULL 4

Range of Values: ( to None)

district:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

3 Blank/Empty/Null/Missing value(s) at index(es): [(29, 'Empty'), (384, 'Empty'), (516, 'Empty')]

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 378 unique value(s):

Category Frequency

Buenos Aires 10

California 9

Shandong 9

West Bengali 9

So Paulo 8

Uttar Pradesh 8

England 7

Maharashtra 7

Southern Tagalog 6

Gois 5

... ...

Wielkopolskie 1

Yerevan 1

Ynlin 1

Zanzibar West 1

Ziguinchor 1

Zufar 1

Zulia 1

al-Daqahliya 1

al-Manama 1

al-Qadarif 1

city\_id (id):

Foreign Key format: All 603 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 600)

References: city.city\_id

postal\_code (postal code):

Postal Code format: Error(s) found:

DQI #1 (Missing Data - Completeness):

4 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'Empty'), (1, 'Empty'), (2, 'Empty'), (3, 'Empty')]

DQI #15 (Domain Violation - Accuracy):

6 Short length alphanumeric value(s) at index(es): [(344, '504'), (397, '966'), (443, '3'), (520, '400'), (551, '841'), (574, '943')]

phone:

Phone format: Error(s) found:

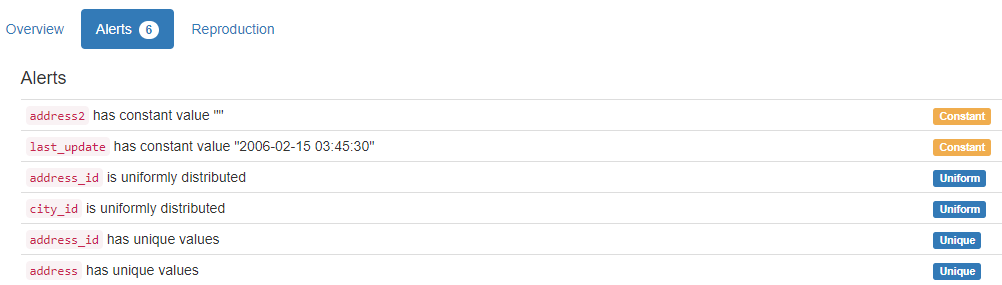
DQI #1 (Missing Data - Completeness):

2 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'Empty'), (1, 'Empty')]

last\_update (timestamp):

Datetime format: All 603 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 03:45 to 2006/02/15 03:45.

YDATA Profiling results: It just found out that the PK column 'address\_id' has unique values and is uniformly distributed. It did not find that column 'address2' contains 599 Empty and 4 NULL values, nor found that column 'district' contains 3 Empty values, 'postal\_code' has 3 Empty values and 'phone' has 2 Empty values :



# DB.table: sakila.city – Retail – Type: Synth

city\_id (id):

Primary Key format: All 600 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 600)

city:

City format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

16 Capitalization/Format issue(s) at index(es): [(11, 'al-Ayn'), (12, 'al-Hawiya'), (13, 'al-Manama'), (14, 'al-Qadarif'), (15, 'al-Qatif'), (189, 'guas Lindas de Gois'), (232, 'Jalib al-Shuyukh'), (362, 'NDjamna'), (384, 'ostka'), (437, 's-Hertogenbosch'), (455, 'Santa Brbara dOeste'), (459, 'Santiago de los Caballeros'), (466, 'Shahr-e Kord'), (475, 'Shubra al-Khayma'), (494, 'Southend-on-Sea'), (561, 'Valle de la Pascua')]

Frequency Distribution (showing top and bottom 10 of 599 cities):

City Frequency

London 2

A Corua (La Corua) 1

Abha 1

Abu Dhabi 1

Acua 1

Adana 1

Addis Abeba 1

Aden 1

Adoni 1

Ahmadnagar 1

... ...

Zhoushan 1

Ziguinchor 1

al-Ayn 1

al-Hawiya 1

al-Manama 1

al-Qadarif 1

al-Qatif 1

guas Lindas de Gois 1

ostka 1

s-Hertogenbosch 1

country\_id (id):

Foreign Key format: All 600 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

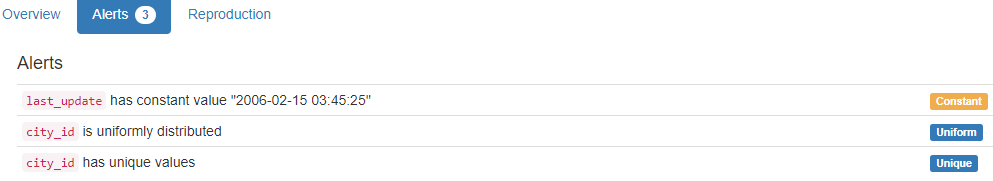
Numeric range of values: (1 : 109)

References: country.country\_id

last\_update (timestamp):

Datetime format: All 600 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 03:45 to 2006/02/15 03:45.

YDATA Profiling results: It just found out that the PK column 'city\_id' has unique values and is uniformly distributed:



# DB.table: sakila.country – Retail – Type: Synth

country\_id (id):

Primary Key format: All 109 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 109)

country:

Country format: All 109 country values are valid.

Frequency Distribution (showing top and bottom 10 of 109 countries):

Country Frequency

Afghanistan 1

Algeria 1

American Samoa 1

Angola 1

Anguilla 1

Argentina 1

Armenia 1

Australia 1

Austria 1

Azerbaijan 1

... ...

Ukraine 1

United Arab Emirates 1

United Kingdom 1

United States 1

Venezuela 1

Vietnam 1

Virgin Islands, U.S. 1

Yemen 1

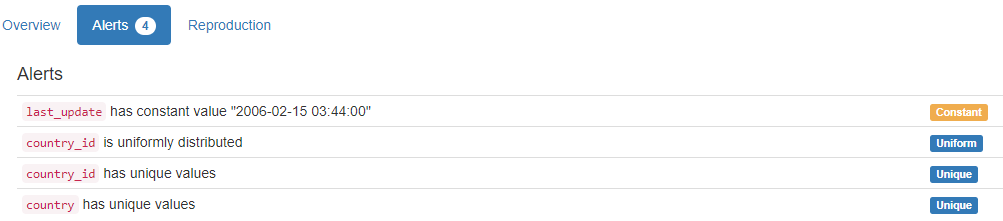
Yugoslavia 1

Zambia 1

last\_update (timestamp):

Datetime format: All 109 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 03:44 to 2006/02/15 03:44.

YDATA Profiling results: It just found out that the PK column 'country\_id' has unique values and is uniformly distributed:



# DB.table: sakila.customer – Retail – Type: Synth

customer\_id (id):

Primary Key format: All 599 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 599)

store\_id (id):

Foreign Key format: All 599 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 2)

References: store.store\_id

first\_name (name):

Name format: All 599 first\_name values are valid.

Frequency Distribution:

first\_name Frequency

JAMIE 2

JESSIE 2

KELLY 2

LESLIE 2

MARION 2

TERRY 2

TRACY 2

WILLIE 2

AARON 1

ADAM 1

... ...

WARREN 1

WAYNE 1

WENDY 1

WESLEY 1

WILLARD 1

WILLIAM 1

WILMA 1

YOLANDA 1

YVONNE 1

ZACHARY 1

Range of Values: (AARON to ZACHARY)

last\_name (name):

Name format: All 599 last\_name values are valid.

Frequency Distribution:

last\_name Frequency

ABNEY 1

ADAM 1

ADAMS 1

ALEXANDER 1

ALLARD 1

ALLEN 1

ALVAREZ 1

ANDERSON 1

ANDREW 1

ANDREWS 1

... ...

WINDHAM 1

WOFFORD 1

WOOD 1

WOODS 1

WREN 1

WRIGHT 1

WYMAN 1

YANEZ 1

YEE 1

YOUNG 1

Range of Values: (ABNEY to YOUNG)

email:

Email format: All 599 email values are valid.

address\_id (id):

Foreign Key format: All 599 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (5 : 605)

References: address.address\_id

active (tinyint):

Numerical format: All 599 values are numerical in the range (0:1).

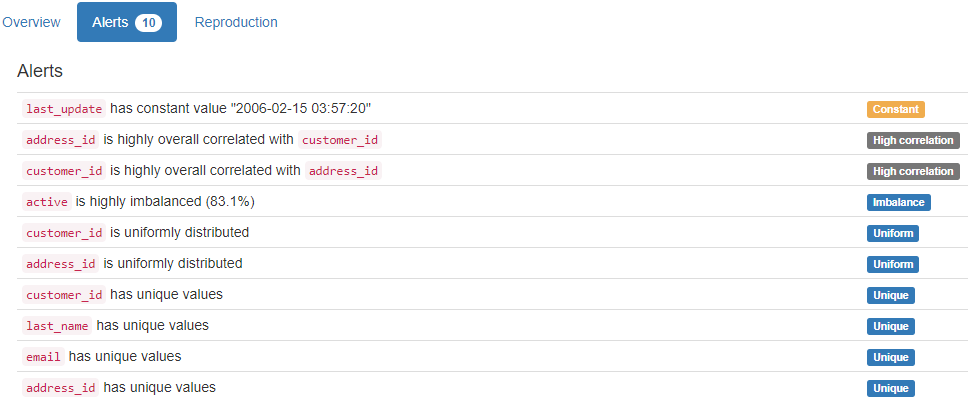
create\_date (datetime):

Datetime format: All 599 datetime values are valid in the YYYYMMDD format in the range 2006/02/14 22:04 to 2006/02/14 22:04.

last\_update (timestamp):

Datetime format: All 599 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 03:57 to 2006/02/15 03:57.

YDATA Profiling results: It just found out that the PK column 'customer\_id' has unique values and is uniformly distributed:



# DB.table: sakila.film – Retail – Type: Synth

film\_id (id):

Primary Key format: All 1000 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 1000)

title:

Categorical format: Error(s) found:

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 1000 unique value(s):

Category Frequency

ACADEMY DINOSAUR 1

ACE GOLDFINGER 1

ADAPTATION HOLES 1

AFFAIR PREJUDICE 1

AFRICAN EGG 1

AGENT TRUMAN 1

AIRPLANE SIERRA 1

AIRPORT POLLOCK 1

ALABAMA DEVIL 1

ALADDIN CALENDAR 1

... ...

WORST BANGER 1

WRATH MILE 1

WRONG BEHAVIOR 1

WYOMING STORM 1

YENTL IDAHO 1

YOUNG LANGUAGE 1

YOUTH KICK 1

ZHIVAGO CORE 1

ZOOLANDER FICTION 1

ZORRO ARK 1

description:

String format: All 1000 string values are valid.

String range (lexicographical): (A Action-Packed Character Study of a Astronaut And a Explorer who must Reach a Monkey in A MySQL Convention : A Unbelieveable Yarn of a Student And a Database Administrator who must Outgun a Husband in An Abandoned Mine Shaft)

release\_year (year):

Year format: All 1000 values are numerical and valid in the range [1800, 2100].

Actual range of values: (2006 : 2006)

language\_id (id):

Foreign Key format: All 1000 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 1)

References: language.language\_id

original\_language\_id (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1000 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (3, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), ('...', '...'), (990, 'NULL'), (991, 'NULL'), (992, 'NULL'), (993, 'NULL'), (994, 'NULL'), (995, 'NULL'), (996, 'NULL'), (997, 'NULL'), (998, 'NULL'), (999, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

No valid values found.

References: language.language\_id

rental\_duration (duration):

Numerical >=0 format: All 1000 values are numerical and greater or equal to 0 in the range (3:7).

rental\_rate (rate):

Numerical format: All 1000 values are numerical in the range (0.99:4.99).

length:

Numerical >=0 format: All 1000 values are numerical and greater or equal to 0 in the range (46:185).

replacement\_cost (cost):

Numerical >=0 format: All 1000 values are numerical and greater or equal to 0 in the range (9.99:29.99).

rating:

Categorical format: All 1000 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

PG-13 223

NC-17 210

R 195

PG 194

G 178

special\_features (features):

Categorical format: All 1000 values are correctly categorical.

Categorical format with 15 unique value(s):

Category Frequency

Trailers,Commentaries,Behind the Scenes 79

Trailers 72

Trailers,Behind the Scenes 72

Trailers,Commentaries 72

Deleted Scenes,Behind the Scenes 71

Behind the Scenes 70

Commentaries,Behind the Scenes 70

Commentaries,Deleted Scenes,Behind the Scenes 66

Trailers,Deleted Scenes 66

Commentaries,Deleted Scenes 65

Trailers,Commentaries,Deleted Scenes 64

Commentaries 62

Deleted Scenes 61

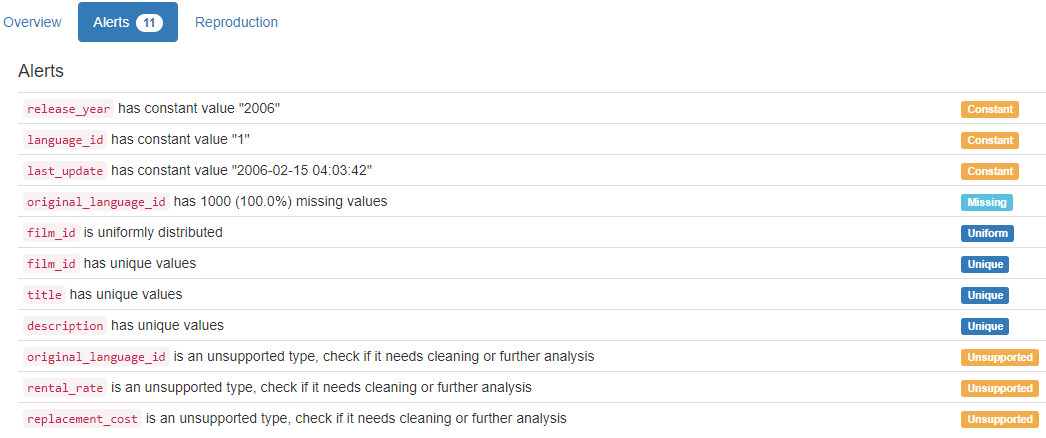
Trailers,Commentaries,Deleted Scenes,Behind the Scenes 61

Trailers,Deleted Scenes,Behind the Scenes 49

last\_update (timestamp):

Datetime format: All 1000 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 04:03 to 2006/02/15 04:03.

YDATA Profiling results: It found the only Missing values column because it contains NULLs. It also found that the PK column 'film\_id' is unique and uniformly distributed.



# DB.table: sakila.inventory – Retail – Type: Synth

inventory\_id (id):

Primary Key format: All 4581 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 4581)

film\_id (id):

Foreign Key format: All 4581 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 1000)

References: film.film\_id

store\_id (id):

Foreign Key format: All 4581 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 2)

References: store.store\_id

last\_update (timestamp):

Datetime format: All 4581 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 04:09 to 2006/02/15 04:09.

YDATA Profiling results: It just found out that the PK column 'inventory\_id' has unique values and is uniformly distributed



# DB.table: sakila.language – Retail – Type: Synth

language\_id (id):

Primary Key format: All 6 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 6)

name:

Name format: All 6 name values are valid.

Frequency Distribution:

name Frequency

English 1

French 1

German 1

Italian 1

Japanese 1

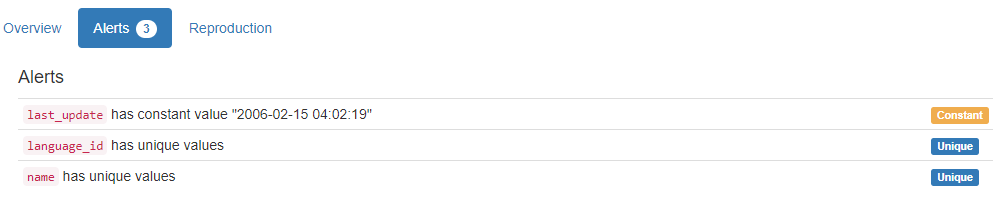
Mandarin 1

Range of Values: (English to Mandarin)

last\_update (timestamp):

Datetime format: All 6 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 04:02 to 2006/02/15 04:02.

YDATA Profiling results: It just found out that the PK column 'language\_id' has unique values:



# DB.table: sakila.rental – Retail – Type: Synth

rental\_id (id):

Primary Key format: All 16044 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 16049)

rental\_date (datetime):

Datetime format: All 16044 datetime values are valid in the YYYYMMDD format in the range 2005/05/24 22:53 to 2006/02/14 15:16.

inventory\_id (id):

Foreign Key format: All 16044 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 4581)

References: inventory.inventory\_id

customer\_id (id):

Foreign Key format: All 16044 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 599)

References: customer.customer\_id

return\_date (datetime):

Datetime format: Error(s) found:

DQI #1 (Missing Data - Completeness):

183 Blank/Empty/Null/Missing value(s) at index(es): [(11491, 'NULL'), (11536, 'NULL'), (11558, 'NULL'), (11572, 'NULL'), (11588, 'NULL'), (11606, 'NULL'), (11641, 'NULL'), (11647, 'NULL'), (11652, 'NULL'), (11667, 'NULL'), ('...', '...'), (15739, 'NULL'), (15772, 'NULL'), (15774, 'NULL'), (15788, 'NULL'), (15807, 'NULL'), (15856, 'NULL'), (15861, 'NULL'), (15869, 'NULL'), (15888, 'NULL'), (15960, 'NULL')] (displaying only the first and last 10 items)

staff\_id (id):

Foreign Key format: All 16044 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 2)

References: staff.staff\_id

last\_update (timestamp):

Datetime format: All 16044 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 20:30 to 2006/02/23 03:12.

YDATA Profiling results: It found the only Missing values column because it contains NULLs. It also found that the PK column 'rental\_id' is unique and uniformly distributed.

A white rectangular object with a black line

Description automatically generated

# DB.table: sakila.staff – Retail – Type: Synth

staff\_id (id):

Primary Key format: All 2 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 2)

first\_name (name):

Name format: All 2 first\_name values are valid.

Frequency Distribution:

first\_name Frequency

Jon 1

Mike 1

Range of Values: (Jon to Mike)

last\_name (name):

Name format: All 2 last\_name values are valid.

Frequency Distribution:

last\_name Frequency

Hillyer 1

Stephens 1

Range of Values: (Hillyer to Stephens)

address\_id (id):

Foreign Key format: All 2 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (3 : 4)

References: address.address\_id

picture:

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

1 Blank/Empty/Null/Missing value(s) at index(es): [(1, 'NULL')]

String range (lexicographical): (BLOB : BLOB)

Frequency Distribution:

picture Frequency

BLOB 1

NULL 1

email:

Email format: All 2 email values are valid.

store\_id (id):

Foreign Key format: All 2 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 2)

References: store.store\_id

active (tinyint):

Numerical format: All 2 values are numerical in the range (1:1).

username:

Name format: All 2 username values are valid.

Frequency Distribution:

username Frequency

Jon 1

Mike 1

Range of Values: (Jon to Mike)

password:

String format: All 2 string values are valid.

String range (lexicographical): (8cb2237d0679ca88db6464eac60da96345513964 : 8cb2237d0679ca88db6464eac60da96345513964)

Frequency Distribution:

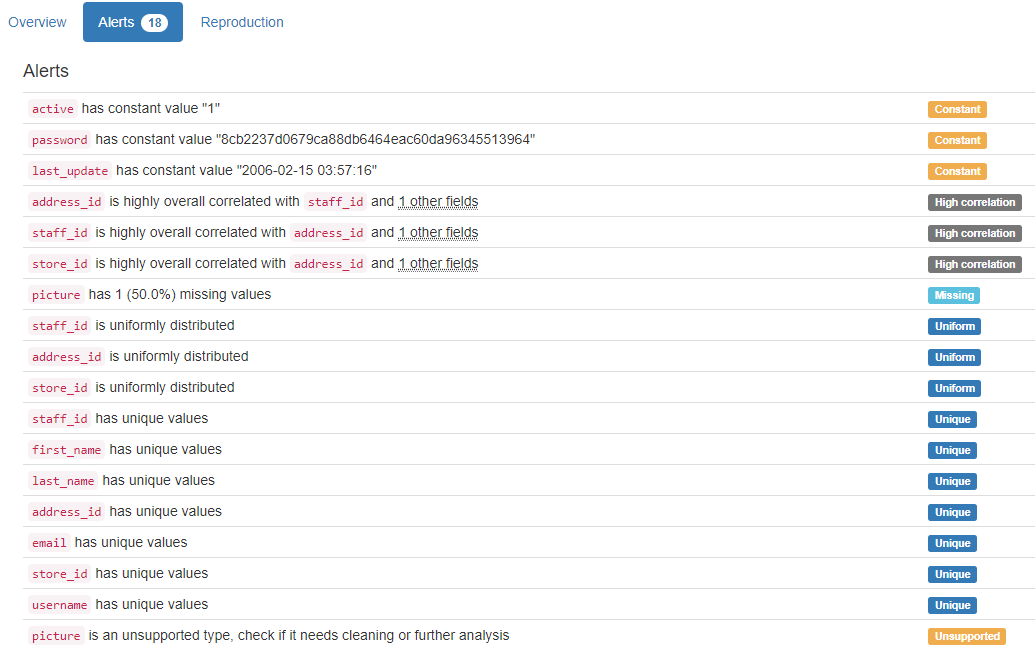
password Frequency

8cb2237d0679ca88db6464eac60da96345513964 2

last\_update (timestamp):

Datetime format: All 2 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 03:57 to 2006/02/15 03:57.

YDATA Profiling results: It found the only Missing values column because it contains 1 NULL value. It also found that the PK column 'staff\_id' is unique and uniformly distributed.



# DB.table: sakila.store – Retail – Type: Synth

store\_id (id):

Primary Key format: All 2 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 2)

manager\_staff\_id (id):

Foreign Key format: All 2 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 2)

References: staff.staff\_id

address\_id (id):

Foreign Key format: All 2 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

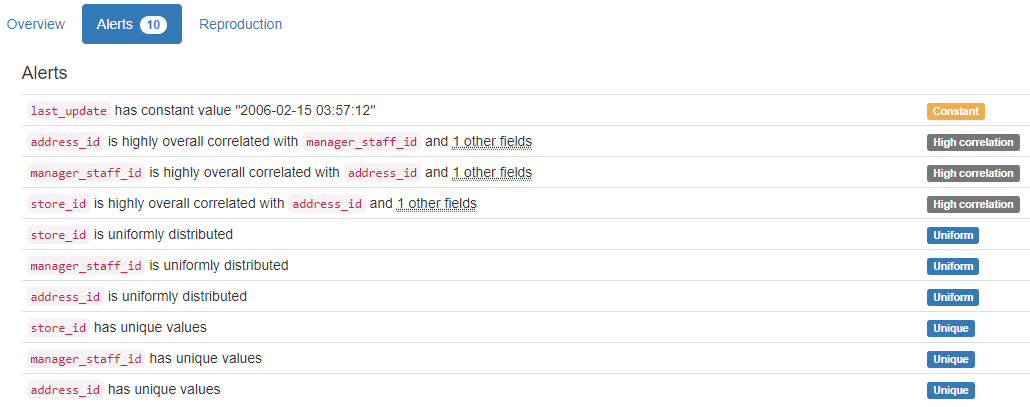
Numeric range of values: (1 : 2)

References: address.address\_id

last\_update (timestamp):

Datetime format: All 2 datetime values are valid in the YYYYMMDD format in the range 2006/02/15 03:57 to 2006/02/15 03:57.

YDATA Profiling results: It just found out that the PK column 'store\_id' has unique values and is uniformly distributed:



# DB.table: SalesDB.Sales – Retail – Type: Synth

Here 6.7 million lines were analysed in less than 3 minutes.

SalesID (id):

Primary Key format: All 6715221 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 6737977)

SalesPersonID (id):

Foreign Key format: All 6715221 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 23)

References: Employees.EmployeeID

CustomerID (id):

Foreign Key format: All 6715221 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 19680)

References: Customers.CustomerID

ProductID (id):

Foreign Key format: All 6715221 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 504)

References: Products.ProductID

Quantity:

Numerical format: All 6715221 values are numerical in the range (1:1042).

YDATA Profiling results: It just found out that the PK column 'SalesID' has unique values:

A white background with a black border

Description automatically generated

# DB.table: stats.users – Education – Type: Real

Id:

Primary Key format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

1 Negative value(s) at index(es): [(0, -1)]

Numeric range of values: (-1 : 55747)

Reputation (integer):

Numerical format: All 40325 values are numerical in the range (1:87393).

CreationDate (datetime):

Datetime format: All 40325 datetime values are valid in the YYYYMMDD format in the range 2010/07/19 06:55 to 2014/09/14 01:01.

DisplayName (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

2 Blank/Empty/Null/Missing value(s) at index(es):

1 NULL values: [(24969, 'NULL')]

1 Non-NULL values: [(2018, 'None')]

DQI #5 (Extraneous Data - Consistency, Uniqueness):

7947 Extraneous data value(s) at index(es): [(28, 'adi92'), (68, 'c4il'), (104, 'd\_2'), (120, 'nbv4'), (162, 'vrish88'), (173, 'user204'), (177, 'user211'), (196, 'm3rLinEz'), (233, 'user283'), (265, 'user333'), ('...', '...'), (40236, 'user21359'), (40237, 'user55659'), (40244, 'user2219907'), (40245, 'user2460415'), (40252, 'user2975986'), (40256, 'user55678'), (40260, 'user2329744'), (40262, 'georgiar18'), (40287, 'rmehta1987'), (40305, 'user56299')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

12989 Capitalization/Format issue(s) at index(es): [(7, 'csgillespie'), (9, 'wahalulu'), (12, 'hannes.koller'), (13, 'slashnick'), (15, 'grokus'), (18, 'radek'), (23, 'balpha'), (25, 'gappy'), (31, 'EAMann'), (36, 'hadley'), ('...', '...'), (40303, 'SuPakarn Junfang'), (40304, 'biggvsdiccvs'), (40307, 'quants\_mum'), (40311, 'jalen'), (40313, 'liszt'), (40314, 'oolusola'), (40315, 'guest'), (40316, 'kpivert'), (40320, 'AussieMeg'), (40322, 'tronbabylove')] (displaying only the first and last 10 items)

Frequency Distribution:

DisplayName Frequency

Chris 80

John 64

Alex 63

Tom 57

Michael 52

James 50

David 48

Dan 44

Matt 43

Andrew 40

... ...

ايمان محمد 1

يآسر الشيبآني 1

ダンボー 1

一二三 1

琼 朱 1

趙維雄 1

陈伟林 1

陳俊宇 1

黃宇陽 1

Ｊ. Ｍ. 1

Range of Values: ( Raid-Mahbouba to Ｊ. Ｍ.)

LastAccessDate (datetime):

Datetime format: All 40325 datetime values are valid in the YYYYMMDD format in the range 2010/07/19 06:55 to 2014/09/14 03:05.

WebsiteUrl (url):

URL format: Error(s) found:

DQI #1 (Missing Data - Completeness):

32204 Blank/Empty/Null/Missing value(s) at index(es): All NULL values:

[(13, 'NULL'), (14, 'NULL'), (17, 'NULL'), (18, 'NULL'), (20, 'NULL'), (22, 'NULL'), (26, 'NULL'), (30, 'NULL'), (37, 'NULL'), (39, 'NULL'), ('...', '...'), (40315, 'NULL'), (40316, 'NULL'), (40317, 'NULL'), (40318, 'NULL'), (40319, 'NULL'), (40320, 'NULL'), (40321, 'NULL'), (40322, 'NULL'), (40323, 'NULL'), (40324, 'NULL')]

DQI #15 (Domain Violation - Accuracy):

356 Invalid URL format issue(s) at index(es): [(51, '[http://None](http://none/)'), (150, '[http://@tzenes](http://tzenes/)'), (160, '[http://none](http://none/)'), (216, '[http://none](http://none/)'), (300, '<http://twitter/simalam>'), (343, '<http://twitter.com/#!/oluies>'), (344, '[http://none](http://none/)'), (356, '[http://N/A](http://n/A)'), (368, '[http://liveatthewitchtrials.blogspot](http://liveatthewitchtrials.blogspot/).'), (441, '[http://somethingelsetodo](http://somethingelsetodo/)'), ('...', '...'), (38403, 'http://-'), (38529, '[http://None](http://none/)'), (38996, '[http://none](http://none/)'), (39130, '[http://nope](http://nope/)'), (39170, '[http://onedoubleapostrophe@gmail.com](http://onedoubleapostrophe@gmail.com/)'), (39374, '[http://cananyoneexplaintomehowabjtransistorworksplease](http://cananyoneexplaintomehowabjtransistorworksplease/)'), (39559, '[http://lovingher@gmail.com](http://lovingher@gmail.com/)'), (39812, '[http://N/A](http://n/A)'), (40111, '[http://none](http://none/)'), (40138, '[http://Missing](http://missing/)')] (displaying only the first and last 10 items)

Location:

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

28643 Blank/Empty/Null/Missing value(s) at index(es):

NULL values (28634): [(14, 'NULL'), (19, 'NULL'), (20, 'NULL'), (22, 'NULL'), (26, 'NULL'), (30, 'NULL'), (37, 'NULL'), (40, 'NULL'), (55, 'NULL'), (58, 'NULL'), ('...', '...'), (40312, 'NULL'), (40313, 'NULL'), (40314, 'NULL'), (40315, 'NULL'), (40316, 'NULL'), (40317, 'NULL'), (40319, 'NULL'), (40320, 'NULL'), (40321, 'NULL'), (40323, 'NULL')]

Non-NULL values (9): [(1111, 'Blank'), (6593, '-'), (8971, '.'), (13628, 'N/A'), (14203, '-'), (16245, 'None'), (17944, '!'), (21566, '-'), (33618, '-')]

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 2465 unique value(s):

Category Frequency

NULL 28634

United States 352

London, United Kingdom 281

Germany 221

United Kingdom 209

India 206

San Francisco, CA 154

New York, United States 148

Seattle, WA 131

Bangalore, India 130

... ...

Đà Nẵng, Việt Nam 1

İstanbul, Turkey 1

Široki Brijeg, Bosnia and Herzegovina 1

Žednik, Serbia 1

Зубин Поток 1

مزرعة يشوع, Lebanon 1

♁ 1

中国 1

仙台 1

北京 1

AboutMe:

Target word not found, and Format not determined

Views:

Numerical >=0 format: All 40325 values are numerical and greater or equal to 0 in the range (0:20932).

UpVotes (integer):

Numerical format: All 40325 values are numerical in the range (0:11442).

DownVotes (integer):

Numerical format: All 40325 values are numerical in the range (0:1920).

AccountId (id):

ID column format: Error(s) found:

DQI #15 (Domain Violation - Accuracy):

1 Negative value(s) at index(es): [(0, -1)]

Numeric range of values: (-1 : 5027354)

Age:

Age format: Error(s) found:

DQI #1 (Missing Data - Completeness):

32007 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (14, 'NULL'), (18, 'NULL'), (20, 'NULL'), (22, 'NULL'), (25, 'NULL'), (26, 'NULL'), (27, 'NULL'), (30, 'NULL'), (35, 'NULL'), ('...', '...'), (40315, 'NULL'), (40316, 'NULL'), (40317, 'NULL'), (40318, 'NULL'), (40319, 'NULL'), (40320, 'NULL'), (40321, 'NULL'), (40322, 'NULL'), (40323, 'NULL'), (40324, 'NULL')] (displaying only the first and last 10 items)

Actual range of values: (13.0 : 94.0)

ProfileImageUrl (url):

URL format: Error(s) found:

DQI #1 (Missing Data - Completeness):

23846 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (1, 'NULL'), (2, 'NULL'), (4, 'NULL'), (5, 'NULL'), (6, 'NULL'), (7, 'NULL'), (8, 'NULL'), (9, 'NULL'), (10, 'NULL'), ('...', '...'), (40296, 'NULL'), (40305, 'NULL'), (40306, 'NULL'), (40307, 'NULL'), (40308, 'NULL'), (40311, 'NULL'), (40314, 'NULL'), (40315, 'NULL'), (40319, 'NULL'), (40321, 'NULL')] (displaying only the first and last 10 items)

DQI #15 (Domain Violation - Accuracy):

14 Invalid URL format issue(s) at index(es): [(1249, '<http://www.gravatar.com/avatar/205ede43beb3ae5dd2695d88e7929612?s=128&amp;d=identicon&amp;r=PG>'), (3278, '<http://www.gravatar.com/avatar/b2d2a783d6bb2ae8c8b4ca9d5237f405?s=128&amp;d=identicon&amp;r=PG>'), (3916, '<http://www.gravatar.com/avatar/f5ef1b81e1aeb3b921a3034d460ce6af?s=128&amp;d=identicon&amp;r=PG>'), (5712, '<http://www.gravatar.com/avatar/a5999e39a83c65b7921361fb82137cd3?s=128&amp;d=identicon&amp;r=PG>'), (7888, '<http://www.gravatar.com/avatar/61c85d77bcaf5212b3a0aa2419c95f33?s=128&amp;d=identicon&amp;r=PG>'), (9007, '<http://www.gravatar.com/avatar/61062bd44817f0d89414856b07a76ce4?s=128&amp;d=identicon&amp;r=PG>'), (14975, '<http://www.gravatar.com/avatar/bd23c25c23fc0c52e8a914fb9d6124c8?s=128&amp;d=identicon&amp;r=PG>'), (17136, '<http://www.gravatar.com/avatar/dd34caad3a10802fa65957709af629b1?s=128&amp;d=identicon&amp;r=PG>'), (17751, '<http://www.gravatar.com/avatar/054dba7842330d7a9ae87d4773847d12?s=128&amp;d=identicon&amp;r=PG>'), (19185, '<http://www.gravatar.com/avatar/6ee30aafb111ccb33f233eb153059bcb?s=128&amp;d=identicon&amp;r=PG>'), (21383, '<http://www.gravatar.com/avatar/363da5c659a30b74d0a74f8eba92e346?s=128&amp;d=identicon&amp;r=PG>'), (27754, '<http://www.gravatar.com/avatar/303b0a741a14439b53c23ce4c9e28a12?s=128&amp;d=identicon&amp;r=PG>'), (37283, '<http://www.gravatar.com/avatar/21d92e60aa7179788093b60c029fe963?s=128&amp;d=identicon&amp;r=PG>'), (38786, '<http://www.gravatar.com/avatar/3d15a12f51296073d9501892423b47ca?s=128&amp;d=identicon&amp;r=PG>')]

YDATA Profiling results: It found correctly 4 of the 6 Missing values columns because they contain NULLs. One of these was the only case where our system did not find, the 'AboutMe' column, which was not associated with any format in our Semantic Type Detection. It did not find all the Missing values for column 'Location'. It just found correctly 13085 cases of 13350 that were found by our system, because there are 265 values with these values: "N/A", "Blank", "!", "-", "." and "None". Besides that it did not find the Missing values on column DisplayName (2: 1 NULL and 1 "None"). It also found that the PK column 'Id' has unique values:

A screenshot of a computer

Description automatically generated

# DB.table: tpcds.item – Retail – Type: Synth

i\_item\_sk (id):

Primary Key format: All 18000 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 18000)

i\_item\_id (id):

ID column format: Error(s) found:

DQI #9 (Duplicates - Uniqueness):

15000 Duplicate value(s) at index(es): [[1, 'AAAAAAAACAAAAAAA'], [2, 'AAAAAAAACAAAAAAA'], [3, 'AAAAAAAAEAAAAAAA'], [4, 'AAAAAAAAEAAAAAAA'], [5, 'AAAAAAAAEAAAAAAA'], [7, 'AAAAAAAAIAAAAAAA'], [8, 'AAAAAAAAIAAAAAAA'], [9, 'AAAAAAAAKAAAAAAA'], [10, 'AAAAAAAAKAAAAAAA'], [11, 'AAAAAAAAKAAAAAAA'], ('...', '...'), [17989, 'AAAAAAAAGEGEAAAA'], [17990, 'AAAAAAAAGEGEAAAA'], [17991, 'AAAAAAAAIEGEAAAA'], [17992, 'AAAAAAAAIEGEAAAA'], [17993, 'AAAAAAAAIEGEAAAA'], [17995, 'AAAAAAAAMEGEAAAA'], [17996, 'AAAAAAAAMEGEAAAA'], [17997, 'AAAAAAAAOEGEAAAA'], [17998, 'AAAAAAAAOEGEAAAA'], [17999, 'AAAAAAAAOEGEAAAA']] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

9000 Uniqueness violation(s) at index(es): [(2, 'AAAAAAAACAAAAAAA'), (4, 'AAAAAAAAEAAAAAAA'), (5, 'AAAAAAAAEAAAAAAA'), (8, 'AAAAAAAAIAAAAAAA'), (10, 'AAAAAAAAKAAAAAAA'), (11, 'AAAAAAAAKAAAAAAA'), (14, 'AAAAAAAAOAAAAAAA'), (16, 'AAAAAAAAABAAAAAA'), (17, 'AAAAAAAAABAAAAAA'), (20, 'AAAAAAAAEBAAAAAA'), ('...', '...'), (17981, 'AAAAAAAAMDGEAAAA'), (17984, 'AAAAAAAAAEGEAAAA'), (17986, 'AAAAAAAACEGEAAAA'), (17987, 'AAAAAAAACEGEAAAA'), (17990, 'AAAAAAAAGEGEAAAA'), (17992, 'AAAAAAAAIEGEAAAA'), (17993, 'AAAAAAAAIEGEAAAA'), (17996, 'AAAAAAAAMEGEAAAA'), (17998, 'AAAAAAAAOEGEAAAA'), (17999, 'AAAAAAAAOEGEAAAA')] (displaying only the first and last 10 items)

Alphanumeric range of values: (AAAAAAAAAAABAAAA : AAAAAAAAPPPBAAAA)

i\_rec\_start\_date (date):

Date format: Error(s) found:

DQI #1 (Missing Data - Completeness):

46 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (1047, 'NULL'), (1328, 'NULL'), (1914, 'NULL'), (2028, 'NULL'), (2722, 'NULL'), (2998, 'NULL'), (3211, 'NULL'), (3349, 'NULL'), (3974, 'NULL'), ('...', '...'), (13554, 'NULL'), (13688, 'NULL'), (13916, 'NULL'), (15190, 'NULL'), (15456, 'NULL'), (15544, 'NULL'), (16196, 'NULL'), (16251, 'NULL'), (17065, 'NULL'), (17391, 'NULL')] (displaying only the first and last 10 items)

Date range: 1997-10-27 to 2001-10-27

i\_rec\_end\_date (date):

Date format: Error(s) found:

DQI #1 (Missing Data - Completeness):

9000 Blank/Empty/Null/Missing value(s) at index(es): [(0, 'NULL'), (2, 'NULL'), (5, 'NULL'), (6, 'NULL'), (8, 'NULL'), (11, 'NULL'), (12, 'NULL'), (14, 'NULL'), (17, 'NULL'), (18, 'NULL'), ('...', '...'), (17981, 'NULL'), (17982, 'NULL'), (17984, 'NULL'), (17987, 'NULL'), (17988, 'NULL'), (17990, 'NULL'), (17993, 'NULL'), (17994, 'NULL'), (17996, 'NULL'), (17999, 'NULL')] (displaying only the first and last 10 items)

Date range: 1999-10-27 to 2001-10-26

i\_item\_desc (item):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

51 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (461, 'NULL'), (779, 'NULL'), (940, 'NULL'), (1047, 'NULL'), (1145, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (1914, 'NULL'), (2243, 'NULL'), ('...', '...'), (15296, 'NULL'), (15456, 'NULL'), (15544, 'NULL'), (16251, 'NULL'), (16400, 'NULL'), (16646, 'NULL'), (16701, 'NULL'), (17065, 'NULL'), (17391, 'NULL'), (17953, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (A : Young, working horses see mentally)

i\_current\_price (price):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

45 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (779, 'NULL'), (940, 'NULL'), (1145, 'NULL'), (1693, 'NULL'), (1914, 'NULL'), (2028, 'NULL'), (2243, 'NULL'), (2998, 'NULL'), (3015, 'NULL'), ('...', '...'), (13178, 'NULL'), (13554, 'NULL'), (14854, 'NULL'), (15253, 'NULL'), (15296, 'NULL'), (15456, 'NULL'), (16400, 'NULL'), (16701, 'NULL'), (17065, 'NULL'), (17391, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.09:99.99).

i\_wholesale\_cost (cost):

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

46 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (940, 'NULL'), (1328, 'NULL'), (1914, 'NULL'), (2028, 'NULL'), (2243, 'NULL'), (2722, 'NULL'), (2998, 'NULL'), (3211, 'NULL'), (3349, 'NULL'), ('...', '...'), (15061, 'NULL'), (15253, 'NULL'), (15296, 'NULL'), (15456, 'NULL'), (16196, 'NULL'), (16251, 'NULL'), (16400, 'NULL'), (16646, 'NULL'), (17391, 'NULL'), (17953, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.02:87.36).

i\_brand\_id (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

44 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (779, 'NULL'), (940, 'NULL'), (1047, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (1914, 'NULL'), (2243, 'NULL'), (2998, 'NULL'), ('...', '...'), (12204, 'NULL'), (13688, 'NULL'), (15190, 'NULL'), (15296, 'NULL'), (15456, 'NULL'), (15544, 'NULL'), (16196, 'NULL'), (16400, 'NULL'), (16646, 'NULL'), (16701, 'NULL')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):

17945 Duplicate value(s) at index(es): [[0.0, 5003002.0], [1.0, 1001001.0], [2.0, 1001001.0], [3.0, 3002001.0], [4.0, 2002002.0], [5.0, 2002002.0], [6.0, 3001002.0], [7.0, 3003001.0], [8.0, 1004002.0], [9.0, 10008011.0], ('...', '...'), [17990.0, 6014007.0], [17991.0, 4002001.0], [17992.0, 4002001.0], [17993.0, 4002001.0], [17994.0, 6003004.0], [17995.0, 8004007.0], [17996.0, 8004007.0], [17997.0, 4004001.0], [17998.0, 4004001.0], [17999.0, 8001001.0]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

17008 Uniqueness violation(s) at index(es): [(2, 1001001.0), (5, 2002002.0), (10, 10008011.0), (11, 10008011.0), (16, 1002001.0), (17, 1002001.0), (22, 10005006.0), (23, 10005006.0), (24, 1004002.0), (26, 2004001.0), ('...', '...'), (17990, 6014007.0), (17991, 4002001.0), (17992, 4002001.0), (17993, 4002001.0), (17994, 6003004.0), (17995, 8004007.0), (17996, 8004007.0), (17997, 4004001.0), (17998, 4004001.0), (17999, 8001001.0)] (displaying only the first and last 10 items)

Numeric range of values: (1001001.0 : 10016017.0)

i\_brand (brand):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

40 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (779, 'NULL'), (1145, 'NULL'), (1328, 'NULL'), (2028, 'NULL'), (2722, 'NULL'), (2998, 'NULL'), (3015, 'NULL'), (4098, 'NULL'), (5029, 'NULL'), ('...', '...'), (13178, 'NULL'), (13424, 'NULL'), (13688, 'NULL'), (13916, 'NULL'), (15190, 'NULL'), (15253, 'NULL'), (15456, 'NULL'), (16196, 'NULL'), (16400, 'NULL'), (17065, 'NULL')]

DQI #10 (Structural Conflicts - Consistency, Uniqueness):

Data seems not categorical or has too many categories (> 100).

Categorical format with 713 unique value(s):

Category Frequency

edu packscholar #1 253

importoscholar #2 253

importoimporto #2 251

exportischolar #1 245

edu packimporto #1 243

edu packscholar #2 243

importoedu pack #2 243

importoimporto #1 242

exportiexporti #1 241

importoedu pack #1 241

... ...

edu packamalgamalg #12 2

edu packamalgamalg #2 2

importoamalgamalg #6 2

namelessunivamalg #2 2

scholaramalgamalg #14 2

scholaramalgamalg #6 2

scholaramalgamalg #7 2

corpamalgamalg #11 1

scholaramalgamalg #5 1

univamalgamalg #12 1

i\_class\_id (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

35 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (461, 'NULL'), (779, 'NULL'), (1047, 'NULL'), (2998, 'NULL'), (3349, 'NULL'), (4641, 'NULL'), (4714, 'NULL'), (4990, 'NULL'), (5563, 'NULL'), ('...', '...'), (13554, 'NULL'), (13688, 'NULL'), (13916, 'NULL'), (15061, 'NULL'), (15296, 'NULL'), (15456, 'NULL'), (16196, 'NULL'), (16251, 'NULL'), (16400, 'NULL'), (16646, 'NULL')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):

17965 Duplicate value(s) at index(es): [[0.0, 3.0], [1.0, 1.0], [2.0, 7.0], [3.0, 2.0], [4.0, 2.0], [5.0, 3.0], [6.0, 1.0], [7.0, 3.0], [8.0, 3.0], [9.0, 8.0], ('...', '...'), [17990.0, 14.0], [17991.0, 2.0], [17992.0, 5.0], [17993.0, 5.0], [17994.0, 3.0], [17995.0, 4.0], [17996.0, 4.0], [17997.0, 4.0], [17998.0, 11.0], [17999.0, 11.0]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

17949 Uniqueness violation(s) at index(es): [(4, 2.0), (5, 3.0), (6, 1.0), (7, 3.0), (8, 3.0), (11, 14.0), (13, 7.0), (14, 2.0), (15, 2.0), (16, 1.0), ('...', '...'), (17990, 14.0), (17991, 2.0), (17992, 5.0), (17993, 5.0), (17994, 3.0), (17995, 4.0), (17996, 4.0), (17997, 4.0), (17998, 11.0), (17999, 11.0)] (displaying only the first and last 10 items)

Numeric range of values: (1.0 : 16.0)

i\_class (class):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

43 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (1047, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (2028, 'NULL'), (2722, 'NULL'), (3349, 'NULL'), (3974, 'NULL'), (4098, 'NULL'), ('...', '...'), (12901, 'NULL'), (13178, 'NULL'), (13688, 'NULL'), (13916, 'NULL'), (14854, 'NULL'), (15190, 'NULL'), (15296, 'NULL'), (16196, 'NULL'), (16646, 'NULL'), (17953, 'NULL')]

Categorical format with 100 unique value(s):

Category Frequency

kids 538

classical 495

shirts 492

mens 483

pop 483

fragrances 465

swimwear 460

toddlers 459

womens 459

athletic 451

... ...

birdal 96

earings 94

blinds/shades 93

business 93

loose stones 93

science 93

arts 90

mystery 90

football 87

NULL 43

i\_category\_id (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

42 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (461, 'NULL'), (779, 'NULL'), (1047, 'NULL'), (1145, 'NULL'), (1914, 'NULL'), (2998, 'NULL'), (3349, 'NULL'), (3408, 'NULL'), (4641, 'NULL'), ('...', '...'), (13554, 'NULL'), (14811, 'NULL'), (14854, 'NULL'), (15190, 'NULL'), (15296, 'NULL'), (15544, 'NULL'), (16646, 'NULL'), (17065, 'NULL'), (17391, 'NULL'), (17953, 'NULL')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):

17958 Duplicate value(s) at index(es): [[0.0, 5.0], [1.0, 1.0], [2.0, 7.0], [3.0, 3.0], [4.0, 2.0], [5.0, 2.0], [6.0, 3.0], [7.0, 3.0], [8.0, 1.0], [9.0, 10.0], ('...', '...'), [17990.0, 3.0], [17991.0, 4.0], [17992.0, 8.0], [17993.0, 4.0], [17994.0, 6.0], [17995.0, 8.0], [17996.0, 1.0], [17997.0, 4.0], [17998.0, 8.0], [17999.0, 8.0]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

17948 Uniqueness violation(s) at index(es): [(5, 2.0), (6, 3.0), (7, 3.0), (8, 1.0), (11, 7.0), (12, 6.0), (14, 5.0), (15, 1.0), (16, 2.0), (17, 1.0), ('...', '...'), (17990, 3.0), (17991, 4.0), (17992, 8.0), (17993, 4.0), (17994, 6.0), (17995, 8.0), (17996, 1.0), (17997, 4.0), (17998, 8.0), (17999, 8.0)] (displaying only the first and last 10 items)

Numeric range of values: (1.0 : 10.0)

i\_category (category):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

43 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (1145, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1914, 'NULL'), (2243, 'NULL'), (2998, 'NULL'), (3211, 'NULL'), (3408, 'NULL'), (4990, 'NULL'), ('...', '...'), (12204, 'NULL'), (12594, 'NULL'), (12901, 'NULL'), (13916, 'NULL'), (15544, 'NULL'), (16251, 'NULL'), (16400, 'NULL'), (16701, 'NULL'), (17391, 'NULL'), (17953, 'NULL')]

Categorical format with 11 unique value(s):

Category Frequency

Music 1860

Shoes 1835

Electronics 1812

Men 1811

Home 1807

Women 1790

Children 1786

Sports 1783

Jewelry 1740

Books 1733

NULL 43

i\_manufact\_id (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

38 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (779, 'NULL'), (940, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (2243, 'NULL'), (2722, 'NULL'), (2998, 'NULL'), (4714, 'NULL'), (4815, 'NULL'), ('...', '...'), (14811, 'NULL'), (14854, 'NULL'), (15253, 'NULL'), (15296, 'NULL'), (15544, 'NULL'), (16196, 'NULL'), (16400, 'NULL'), (16646, 'NULL'), (16701, 'NULL'), (17391, 'NULL')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):

17945 Duplicate value(s) at index(es): [[0.0, 52.0], [1.0, 294.0], [2.0, 294.0], [3.0, 479.0], [4.0, 220.0], [5.0, 212.0], [6.0, 214.0], [7.0, 630.0], [8.0, 630.0], [9.0, 350.0], ('...', '...'), [17990.0, 649.0], [17991.0, 356.0], [17992.0, 515.0], [17993.0, 515.0], [17994.0, 460.0], [17995.0, 225.0], [17996.0, 629.0], [17997.0, 172.0], [17998.0, 191.0], [17999.0, 191.0]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

16968 Uniqueness violation(s) at index(es): [(2, 294.0), (8, 630.0), (16, 117.0), (17, 117.0), (20, 995.0), (23, 260.0), (30, 167.0), (34, 553.0), (35, 553.0), (50, 140.0), ('...', '...'), (17990, 649.0), (17991, 356.0), (17992, 515.0), (17993, 515.0), (17994, 460.0), (17995, 225.0), (17996, 629.0), (17997, 172.0), (17998, 191.0), (17999, 191.0)] (displaying only the first and last 10 items)

Numeric range of values: (1.0 : 1000.0)

i\_manufact (char):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

43 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (2243, 'NULL'), (2722, 'NULL'), (3015, 'NULL'), (3349, 'NULL'), (3408, 'NULL'), (3974, 'NULL'), ('...', '...'), (13424, 'NULL'), (13916, 'NULL'), (14811, 'NULL'), (15061, 'NULL'), (15190, 'NULL'), (15253, 'NULL'), (15296, 'NULL'), (15544, 'NULL'), (16196, 'NULL'), (16400, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (able : pripripri)

Frequency Distribution:

i\_manufact Frequency

callyoughtought 58

ationeseought 56

eingantiought 55

prioughtpri 54

callypri 50

antin stable 49

ationbarable 48

eingese 48

baroughtought 47

oughtoughtpri 47

... ...

callyantin st 1

eingeingn st 1

eingoughtn st 1

esebareing 1

eseoughtn st 1

n stoughteing 1

oughtbarn st 1

oughtpriation 1

priableeing 1

priablen st 1

i\_size (size):

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

11020 Blank/Empty/Null/Missing value(s) at index(es): [(0, '"N/A"'), (2, '"N/A"'), (9, '"N/A"'), (10, '"N/A"'), (11, '"N/A"'), (12, '"N/A"'), (13, '"N/A"'), (14, '"N/A"'), (20, '"N/A"'), (21, '"N/A"'), ('...', '...'), (17984, '"N/A"'), (17985, '"N/A"'), (17986, '"N/A"'), (17988, '"N/A"'), (17989, '"N/A"'), (17992, '"N/A"'), (17994, '"N/A"'), (17995, '"N/A"'), (17998, '"N/A"'), (17999, '"N/A"')] (displaying only the first and last 10 items)

DQI #17 (Wrong Data Type - Consistency):

6980 Non-numeric value(s) at index(es): [(1, 'petite'), (3, 'extra large'), (4, 'petite'), (5, 'large'), (6, 'petite'), (7, 'extra large'), (8, 'medium'), (15, 'large'), (16, 'extra large'), (17, 'medium'), ('...', '...'), (17963, 'medium'), (17972, 'medium'), (17974, 'large'), (17977, 'medium'), (17987, 'medium'), (17990, 'petite'), (17991, 'medium'), (17993, 'extra large'), (17996, 'petite'), (17997, 'large')] (displaying only the first and last 10 items)

No valid numerical values found.

i\_formulation (formulation):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

43 Blank/Empty/Null/Missing value(s) at index(es): [(180, 'NULL'), (940, 'NULL'), (1328, 'NULL'), (1693, 'NULL'), (1914, 'NULL'), (2028, 'NULL'), (3974, 'NULL'), (4098, 'NULL'), (4641, 'NULL'), (5146, 'NULL'), ('...', '...'), (14854, 'NULL'), (15061, 'NULL'), (15190, 'NULL'), (15296, 'NULL'), (15456, 'NULL'), (15544, 'NULL'), (16251, 'NULL'), (16701, 'NULL'), (17391, 'NULL'), (17953, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (0001145946rose283864 : yellow97669389506333)

Frequency Distribution:

i\_formulation Frequency

NULL 43

00588802055807peach8 3

0064sky6890999816999 3

007smoke758094520928 3

008217751thistle4373 3

00turquoise939138765 3

01066tan734187726993 3

014340058goldenrod58 3

01469987178440sky790 3

01537139789sky451400 3

... ...

yellow57856820939171 1

yellow58212992535573 1

yellow59211779642537 1

yellow71395936909612 1

yellow84066692212763 1

yellow89829491246347 1

yellow90447611874549 1

yellow91077138175627 1

yellow92939674830318 1

yellow97669389506333 1

i\_color (color):

Categorical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

45 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (779, 'NULL'), (940, 'NULL'), (1047, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (2243, 'NULL'), (2722, 'NULL'), (2998, 'NULL'), ('...', '...'), (13916, 'NULL'), (14811, 'NULL'), (14854, 'NULL'), (15061, 'NULL'), (15190, 'NULL'), (15253, 'NULL'), (16251, 'NULL'), (16646, 'NULL'), (16701, 'NULL'), (17953, 'NULL')]

Categorical format with 93 unique value(s):

Category Frequency

turquoise 449

slate 424

smoke 424

thistle 410

peach 409

spring 407

saddle 406

salmon 405

tan 405

papaya 402

... ...

burlywood 38

burnished 38

chocolate 38

cornflower 38

frosted 38

azure 35

dark 34

firebrick 34

coral 30

cream 30

i\_units (units):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

925 Blank/Empty/Null/Missing value(s) at index(es): [(44, '"N/A"'), (91, '"N/A"'), (99, '"N/A"'), (101, '"N/A"'), (135, '"N/A"'), (143, '"N/A"'), (164, '"N/A"'), (180, 'NULL'), (185, '"N/A"'), (191, '"N/A"'), ('...', '...'), (17786, '"N/A"'), (17862, '"N/A"'), (17868, '"N/A"'), (17892, '"N/A"'), (17903, '"N/A"'), (17918, '"N/A"'), (17940, '"N/A"'), (17955, '"N/A"'), (17958, '"N/A"'), (17984, '"N/A"')] (displaying only the first and last 10 items)

DQI #17 (Wrong Data Type - Consistency):

17075 Non-numeric value(s) at index(es): [(0, 'Tsp'), (1, 'Bunch'), (2, 'Cup'), (3, 'Tbl'), (4, 'Cup'), (5, 'Bundle'), (6, 'Tsp'), (7, 'Carton'), (8, 'Tbl'), (9, 'Tsp'), ('...', '...'), (17990, 'Bundle'), (17991, 'Ounce'), (17992, 'Cup'), (17993, 'Box'), (17994, 'Dozen'), (17995, 'Dozen'), (17996, 'Bunch'), (17997, 'Ounce'), (17998, 'Ton'), (17999, 'Ton')] (displaying only the first and last 10 items)

No valid numeric values found.

i\_container (container):

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

39 Blank/Empty/Null/Missing value(s) at index(es): [(940, 'NULL'), (1047, 'NULL'), (1145, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1693, 'NULL'), (2243, 'NULL'), (3349, 'NULL'), (3408, 'NULL'), (3974, 'NULL'), ('...', '...'), (11629, 'NULL'), (11891, 'NULL'), (12594, 'NULL'), (12901, 'NULL'), (13178, 'NULL'), (14854, 'NULL'), (15061, 'NULL'), (16646, 'NULL'), (17065, 'NULL'), (17953, 'NULL')] (displaying only the first and last 10 items)

String range (lexicographical): (Unknown : Unknown)

Frequency Distribution:

i\_container Frequency

Unknown 17961

NULL 39

i\_manager\_id (id):

ID column format: Error(s) found:

DQI #1 (Missing Data - Completeness):

41 Blank/Empty/Null/Missing value(s) at index(es): [(1145, 'NULL'), (1328, 'NULL'), (1482, 'NULL'), (1914, 'NULL'), (2243, 'NULL'), (2722, 'NULL'), (3015, 'NULL'), (3974, 'NULL'), (4815, 'NULL'), (4990, 'NULL'), ('...', '...'), (11891, 'NULL'), (12901, 'NULL'), (15061, 'NULL'), (15456, 'NULL'), (15544, 'NULL'), (16400, 'NULL'), (16701, 'NULL'), (17065, 'NULL'), (17391, 'NULL'), (17953, 'NULL')] (displaying only the first and last 10 items)

DQI #9 (Duplicates - Uniqueness):

17959 Duplicate value(s) at index(es): [[0.0, 6.0], [1.0, 98.0], [2.0, 18.0], [3.0, 26.0], [4.0, 27.0], [5.0, 6.0], [6.0, 64.0], [7.0, 25.0], [8.0, 3.0], [9.0, 34.0], ('...', '...'), [17990.0, 65.0], [17991.0, 5.0], [17992.0, 35.0], [17993.0, 22.0], [17994.0, 24.0], [17995.0, 18.0], [17996.0, 61.0], [17997.0, 20.0], [17998.0, 32.0], [17999.0, 1.0]] (displaying only the first and last 10 items)

DQI #19 (Uniqueness Violation - Uniqueness):

17859 Uniqueness violation(s) at index(es): [(5, 6.0), (10, 26.0), (13, 6.0), (24, 18.0), (26, 31.0), (29, 32.0), (30, 98.0), (32, 51.0), (38, 34.0), (39, 64.0), ('...', '...'), (17990, 65.0), (17991, 5.0), (17992, 35.0), (17993, 22.0), (17994, 24.0), (17995, 18.0), (17996, 61.0), (17997, 20.0), (17998, 32.0), (17999, 1.0)] (displaying only the first and last 10 items)

Numeric range of values: (1.0 : 100.0)

i\_product\_name (name):

Name format: Error(s) found:

DQI #1 (Missing Data - Completeness):

36 Blank/Empty/Null/Missing value(s) at index(es): [(461, 'NULL'), (779, 'NULL'), (1328, 'NULL'), (2028, 'NULL'), (2243, 'NULL'), (2722, 'NULL'), (2998, 'NULL'), (3015, 'NULL'), (3211, 'NULL'), (4098, 'NULL'), ('...', '...'), (11515, 'NULL'), (11629, 'NULL'), (12055, 'NULL'), (12594, 'NULL'), (12901, 'NULL'), (15253, 'NULL'), (15296, 'NULL'), (15544, 'NULL'), (16400, 'NULL'), (16701, 'NULL')] (displaying only the first and last 10 items)

Frequency Distribution:

i\_product\_name Frequency

NULL 36

able 1

ableable 1

ableableable 1

ableableableable 1

ableableableableought 1

ableableableanti 1

ableableableantiought 1

ableableableation 1

ableableableationought 1

... ...

pripriprically 1

pripripricallyought 1

pripriprieing 1

pripripriese 1

pripriprieseought 1

pripriprin st 1

pripripriought 1

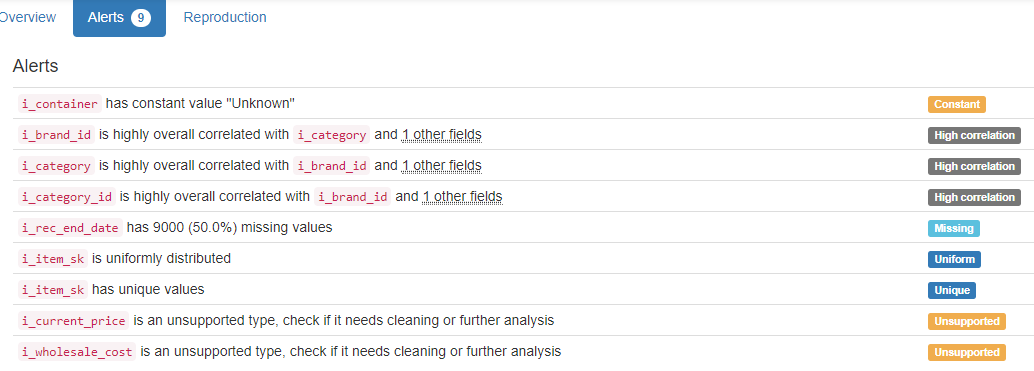
pripriprioughtought 1

pripripripri 1

pripripripriought 1

Range of Values: (able to pripripripriought)

YDATA Profiling results: It found only 1 ('i\_rec\_end\_date') out of the 20 Missing values columns because it contains NULLs. However, it did not find all the Missing values for columns 'i\_rec\_start\_date', 'i\_item\_desc', 'i\_current\_price', 'i\_wholesale\_cost', 'i\_brand\_id', 'i\_brand', 'i\_class\_id', 'i\_class', 'i\_category\_id', 'i\_category', 'i\_manufact\_id', 'i\_manufact', 'i\_formulation', 'i\_color', 'i\_container', 'i\_manager\_id', 'i\_product\_name' that were all NULLs. It also did not find the Missing values of columns 'i\_units' that contains "NULL" and "N/A" and 'i\_size' that contains only "N/A".It also found that the PK column 'i\_item\_sk' is Unique and Uniformly distributed.



# DB.table: tpcds.store\_sales – Retail – Type: Synth

Almost 3 million lines are analysed in this database table and 23 columns!

ss\_item\_sk|ss\_ticket\_number (Composite Primary Key):

Composite Primary Key format: All 2880404 ID values are unique and valid, and thus suitable for use as a Primary Key.

Alphanumeric range of values: (10000|10035 : 9|99921)

ss\_sold\_date\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130093 Blank/Empty/Null/Missing value(s) at index(es): [(11, 'NULL'), (13, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (134, 'NULL'), (163, 'NULL'), (199, 'NULL'), (206, 'NULL'), ('...', '...'), (2880198, 'NULL'), (2880233, 'NULL'), (2880239, 'NULL'), (2880259, 'NULL'), (2880314, 'NULL'), (2880371, 'NULL'), (2880389, 'NULL'), (2880394, 'NULL'), (2880397, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (2450816.0 : 2452642.0)

References: date\_dim.d\_date\_sk

ss\_sold\_time\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129637 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (38, 'NULL'), (64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (130, 'NULL'), (163, 'NULL'), (199, 'NULL'), (208, 'NULL'), (217, 'NULL'), ('...', '...'), (2880239, 'NULL'), (2880259, 'NULL'), (2880278, 'NULL'), (2880295, 'NULL'), (2880307, 'NULL'), (2880357, 'NULL'), (2880378, 'NULL'), (2880389, 'NULL'), (2880394, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (28800.0 : 75599.0)

References: time\_dim.t\_time\_sk

ss\_customer\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129752 Blank/Empty/Null/Missing value(s) at index(es): [(13, 'NULL'), (47, 'NULL'), (64, 'NULL'), (78, 'NULL'), (92, 'NULL'), (97, 'NULL'), (134, 'NULL'), (163, 'NULL'), (188, 'NULL'), (206, 'NULL'), ('...', '...'), (2880244, 'NULL'), (2880257, 'NULL'), (2880259, 'NULL'), (2880307, 'NULL'), (2880314, 'NULL'), (2880347, 'NULL'), (2880357, 'NULL'), (2880371, 'NULL'), (2880394, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (1.0 : 100000.0)

References: customer.c\_customer\_sk

ss\_cdemo\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129700 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (11, 'NULL'), (13, 'NULL'), (47, 'NULL'), (64, 'NULL'), (78, 'NULL'), (134, 'NULL'), (206, 'NULL'), (217, 'NULL'), (222, 'NULL'), ('...', '...'), (2880180, 'NULL'), (2880198, 'NULL'), (2880233, 'NULL'), (2880239, 'NULL'), (2880257, 'NULL'), (2880347, 'NULL'), (2880371, 'NULL'), (2880389, 'NULL'), (2880394, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (15.0 : 1920797.0)

References: customer\_demographics.cd\_demo\_sk

ss\_hdemo\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129847 Blank/Empty/Null/Missing value(s) at index(es): [(47, 'NULL'), (64, 'NULL'), (130, 'NULL'), (134, 'NULL'), (208, 'NULL'), (232, 'NULL'), (256, 'NULL'), (280, 'NULL'), (381, 'NULL'), (431, 'NULL'), ('...', '...'), (2880244, 'NULL'), (2880295, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880371, 'NULL'), (2880389, 'NULL'), (2880390, 'NULL'), (2880394, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (1.0 : 7200.0)

References: household\_demographics.hd\_demo\_sk

ss\_addr\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129975 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (134, 'NULL'), (206, 'NULL'), (228, 'NULL'), (256, 'NULL'), ('...', '...'), (2880239, 'NULL'), (2880257, 'NULL'), (2880278, 'NULL'), (2880307, 'NULL'), (2880314, 'NULL'), (2880357, 'NULL'), (2880371, 'NULL'), (2880378, 'NULL'), (2880389, 'NULL'), (2880397, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (1.0 : 50000.0)

References: customer\_address.ca\_address\_sk

ss\_store\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130034 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (13, 'NULL'), (47, 'NULL'), (57, 'NULL'), (64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (163, 'NULL'), ('...', '...'), (2880191, 'NULL'), (2880198, 'NULL'), (2880239, 'NULL'), (2880259, 'NULL'), (2880307, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880378, 'NULL'), (2880394, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (1.0 : 10.0)

References: store.s\_store\_sk

ss\_promo\_sk (id):

Foreign Key format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129484 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (13, 'NULL'), (47, 'NULL'), (64, 'NULL'), (92, 'NULL'), (130, 'NULL'), (163, 'NULL'), (199, 'NULL'), (206, 'NULL'), (208, 'NULL'), ('...', '...'), (2880295, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880357, 'NULL'), (2880371, 'NULL'), (2880377, 'NULL'), (2880378, 'NULL'), (2880389, 'NULL'), (2880390, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Note: This is a Foreign Key, so duplicate values are allowed.

Numeric range of values: (1.0 : 300.0)

References: promotion.p\_promo\_sk

ss\_quantity (quantity):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129996 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (11, 'NULL'), (13, 'NULL'), (40, 'NULL'), (71, 'NULL'), (78, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (163, 'NULL'), ('...', '...'), (2880239, 'NULL'), (2880244, 'NULL'), (2880259, 'NULL'), (2880278, 'NULL'), (2880295, 'NULL'), (2880347, 'NULL'), (2880357, 'NULL'), (2880377, 'NULL'), (2880389, 'NULL'), (2880390, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:100.0).

ss\_wholesale\_cost (cost):

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130023 Blank/Empty/Null/Missing value(s) at index(es): [(38, 'NULL'), (40, 'NULL'), (57, 'NULL'), (71, 'NULL'), (206, 'NULL'), (208, 'NULL'), (217, 'NULL'), (222, 'NULL'), (228, 'NULL'), (256, 'NULL'), ('...', '...'), (2880244, 'NULL'), (2880257, 'NULL'), (2880278, 'NULL'), (2880307, 'NULL'), (2880314, 'NULL'), (2880357, 'NULL'), (2880371, 'NULL'), (2880389, 'NULL'), (2880394, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:100.0).

ss\_list\_price (price):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130003 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (13, 'NULL'), (38, 'NULL'), (47, 'NULL'), (64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (134, 'NULL'), ('...', '...'), (2880244, 'NULL'), (2880259, 'NULL'), (2880278, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880357, 'NULL'), (2880390, 'NULL'), (2880397, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:200.0).

ss\_sales\_price (price):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129666 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (11, 'NULL'), (38, 'NULL'), (64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (188, 'NULL'), (199, 'NULL'), ('...', '...'), (2880180, 'NULL'), (2880278, 'NULL'), (2880295, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880357, 'NULL'), (2880377, 'NULL'), (2880389, 'NULL'), (2880394, 'NULL'), (2880397, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:199.56).

ss\_ext\_discount\_amt (decimal):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129838 Blank/Empty/Null/Missing value(s) at index(es): [(11, 'NULL'), (13, 'NULL'), (38, 'NULL'), (40, 'NULL'), (47, 'NULL'), (57, 'NULL'), (71, 'NULL'), (130, 'NULL'), (134, 'NULL'), (163, 'NULL'), ('...', '...'), (2880160, 'NULL'), (2880180, 'NULL'), (2880197, 'NULL'), (2880239, 'NULL'), (2880257, 'NULL'), (2880278, 'NULL'), (2880295, 'NULL'), (2880339, 'NULL'), (2880378, 'NULL'), (2880389, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:17588.25).

ss\_ext\_sales\_price (price):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130327 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (38, 'NULL'), (47, 'NULL'), (57, 'NULL'), (64, 'NULL'), (97, 'NULL'), (134, 'NULL'), (163, 'NULL'), (206, 'NULL'), (222, 'NULL'), ('...', '...'), (2880233, 'NULL'), (2880257, 'NULL'), (2880278, 'NULL'), (2880307, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880357, 'NULL'), (2880397, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:19562.4).

ss\_ext\_wholesale\_cost (cost):

Numerical >=0 format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130044 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (11, 'NULL'), (13, 'NULL'), (40, 'NULL'), (47, 'NULL'), (57, 'NULL'), (64, 'NULL'), (71, 'NULL'), (78, 'NULL'), (92, 'NULL'), ('...', '...'), (2880191, 'NULL'), (2880194, 'NULL'), (2880239, 'NULL'), (2880244, 'NULL'), (2880257, 'NULL'), (2880347, 'NULL'), (2880371, 'NULL'), (2880378, 'NULL'), (2880390, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.0:10000.0).

ss\_ext\_list\_price (price):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129933 Blank/Empty/Null/Missing value(s) at index(es): [(13, 'NULL'), (38, 'NULL'), (47, 'NULL'), (57, 'NULL'), (64, 'NULL'), (78, 'NULL'), (92, 'NULL'), (130, 'NULL'), (163, 'NULL'), (188, 'NULL'), ('...', '...'), (2880295, 'NULL'), (2880307, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880357, 'NULL'), (2880371, 'NULL'), (2880377, 'NULL'), (2880389, 'NULL'), (2880390, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (1.12:19984.0).

ss\_ext\_tax (tax):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130410 Blank/Empty/Null/Missing value(s) at index(es): [(64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (134, 'NULL'), (188, 'NULL'), (217, 'NULL'), (222, 'NULL'), (228, 'NULL'), (323, 'NULL'), ('...', '...'), (2880113, 'NULL'), (2880160, 'NULL'), (2880180, 'NULL'), (2880194, 'NULL'), (2880257, 'NULL'), (2880307, 'NULL'), (2880347, 'NULL'), (2880371, 'NULL'), (2880377, 'NULL'), (2880397, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:1749.87).

ss\_coupon\_amt (decimal):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129838 Blank/Empty/Null/Missing value(s) at index(es): [(11, 'NULL'), (13, 'NULL'), (38, 'NULL'), (40, 'NULL'), (47, 'NULL'), (57, 'NULL'), (71, 'NULL'), (130, 'NULL'), (134, 'NULL'), (163, 'NULL'), ('...', '...'), (2880160, 'NULL'), (2880180, 'NULL'), (2880197, 'NULL'), (2880239, 'NULL'), (2880257, 'NULL'), (2880278, 'NULL'), (2880295, 'NULL'), (2880339, 'NULL'), (2880378, 'NULL'), (2880389, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:17588.25).

ss\_net\_paid (paid):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

129397 Blank/Empty/Null/Missing value(s) at index(es): [(11, 'NULL'), (47, 'NULL'), (57, 'NULL'), (64, 'NULL'), (71, 'NULL'), (78, 'NULL'), (92, 'NULL'), (130, 'NULL'), (188, 'NULL'), (206, 'NULL'), ('...', '...'), (2880314, 'NULL'), (2880339, 'NULL'), (2880357, 'NULL'), (2880371, 'NULL'), (2880377, 'NULL'), (2880378, 'NULL'), (2880389, 'NULL'), (2880394, 'NULL'), (2880397, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:19562.4).

ss\_net\_paid\_inc\_tax (paid):

Numerical format: Error(s) found:

DQI #1 (Missing Data - Completeness):

130022 Blank/Empty/Null/Missing value(s) at index(es): [(13, 'NULL'), (47, 'NULL'), (57, 'NULL'), (64, 'NULL'), (71, 'NULL'), (92, 'NULL'), (97, 'NULL'), (130, 'NULL'), (134, 'NULL'), (188, 'NULL'), ('...', '...'), (2880278, 'NULL'), (2880295, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880357, 'NULL'), (2880390, 'NULL'), (2880394, 'NULL'), (2880397, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (0.0:21192.87).

ss\_net\_profit (profit):

Numerical format: Error(s) found:

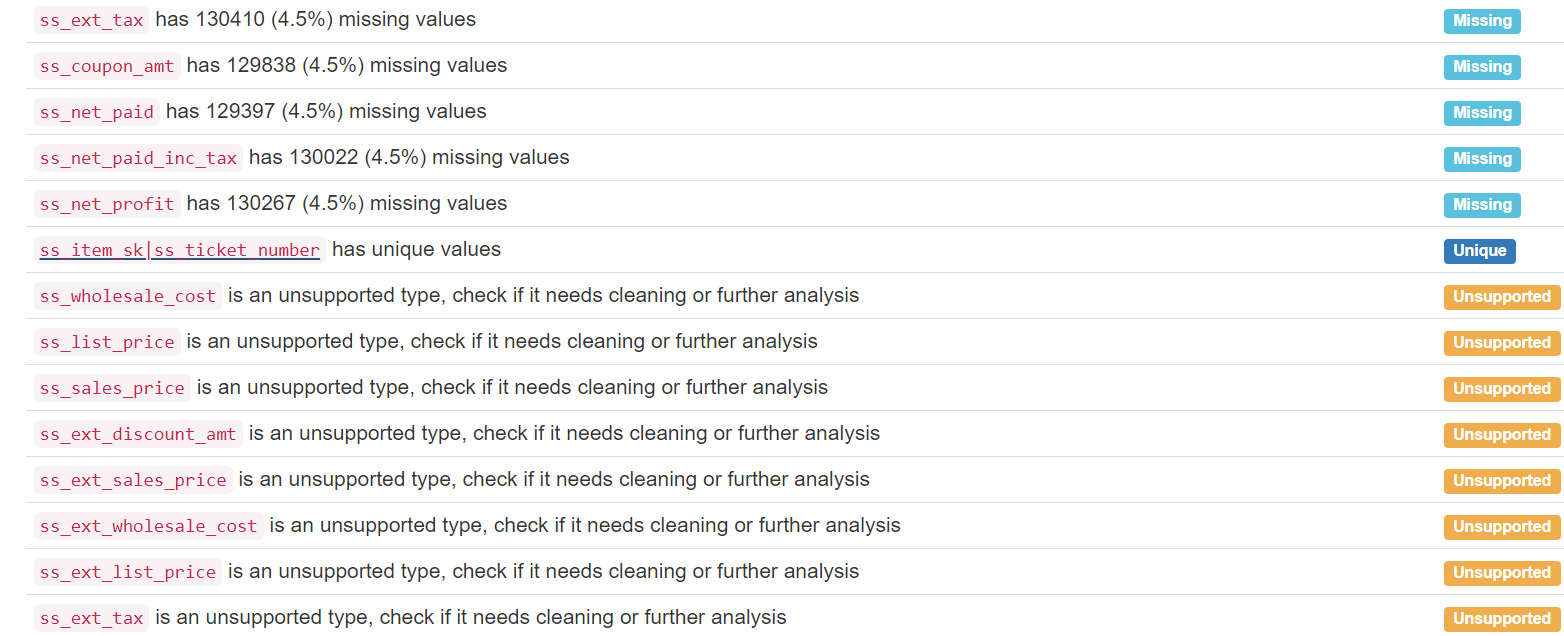
DQI #1 (Missing Data - Completeness):

130267 Blank/Empty/Null/Missing value(s) at index(es): [(7, 'NULL'), (38, 'NULL'), (40, 'NULL'), (71, 'NULL'), (78, 'NULL'), (134, 'NULL'), (199, 'NULL'), (232, 'NULL'), (256, 'NULL'), (258, 'NULL'), ('...', '...'), (2880259, 'NULL'), (2880295, 'NULL'), (2880314, 'NULL'), (2880339, 'NULL'), (2880347, 'NULL'), (2880377, 'NULL'), (2880390, 'NULL'), (2880394, 'NULL'), (2880397, 'NULL'), (2880403, 'NULL')] (displaying only the first and last 10 items)

Range of values: (-9969.53:9731.7).

YDATA Profiling results: It found the same 21 Missing values columns because they contain NULLs. It also found that the Composite PK column 'ss\_item\_sk|ss\_ticket\_number' is unique.





A screenshot of a phone

Description automatically generated

# DB.table: trains.cars – Logistic – Type: Synth

id:

Primary Key format: All 63 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (1 : 63)

train\_id (id):

Foreign Key format: All 63 ID values are valid for use as a Foreign Key, and duplicate values are allowed.

Numeric range of values: (1 : 20)

References: trains.id

position:

Categorical format: All 63 values are correctly categorical.

Categorical format with 4 unique value(s):

Category Frequency

1 20

2 20

3 14

4 9

shape:

Categorical format: All 63 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

rectangle 44

u\_shaped 9

bucket 7

hexagon 2

ellipse 1

len (length):

Numerical >=0 format: Error(s) found:

DQI #17 (Wrong Data Type - Consistency):

63 Non-numeric value(s) at index(es): [(0, 'short'), (1, 'long'), (2, 'short'), (3, 'long'), (4, 'short'), (5, 'short'), (6, 'short'), (7, 'long'), (8, 'short'), (9, 'short'), ('...', '...'), (53, 'short'), (54, 'long'), (55, 'short'), (56, 'long'), (57, 'long'), (58, 'long'), (59, 'long'), (60, 'short'), (61, 'long'), (62, 'short')] (displaying only the first and last 10 items)

No valid numerical values found.

sides (side):

String format: All 63 string values are valid.

String range (lexicographical): (double : not\_double)

roof:

String format: Error(s) found:

DQI #1 (Missing Data - Completeness):

41 Blank/Empty/Null/Missing value(s) at index(es): [(0, '"none"'), (1, '"none"'), (3, '"none"'), (5, '"none"'), (6, '"none"'), (9, '"none"'), (10, '"none"'), (12, '"none"'), (13, '"none"'), (16, '"none"'), ('...', '...'), (49, '"none"'), (51, '"none"'), (52, '"none"'), (53, '"none"'), (54, '"none"'), (55, '"none"'), (56, '"none"'), (59, '"none"'), (60, '"none"'), (62, '"none"')] (displaying only the first and last 10 items)

String range (lexicographical): (arc : peaked)

wheels (integer):

Numerical format: All 63 values are numerical in the range (2:3).

load\_shape (shape):

Categorical format: All 63 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

rectangle 22

triangle 19

circle 16

hexagon 5

diamond 1

load\_num (num):

Numerical format: All 63 values are numerical in the range (0:3).

YDATA Profiling results: It just found out that the PK column 'id' has unique values and is uniformly distributed:

A screenshot of a computer

Description automatically generated

# DB.table: UW\_std.person – Education – Type: Real

p\_id (id):

Primary Key format: All 278 ID values are unique and valid, and thus suitable for use as a Primary Key.

Numeric range of values: (3 : 435)

professor:

Name format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

278 Extraneous data value(s) at index(es): [(0, '0'), (1, '0'), (2, '1'), (3, '0'), (4, '1'), (5, '0'), (6, '0'), (7, '0'), (8, '0'), (9, '0'), ('...', '...'), (268, '0'), (269, '0'), (270, '0'), (271, '0'), (272, '0'), (273, '0'), (274, '0'), (275, '0'), (276, '0'), (277, '0')] (displaying only the first and last 10 items)

Frequency Distribution:

professor Frequency

0 216

1 62

Range of Values: (0 to 1)

student:

Name format: Error(s) found:

DQI #5 (Extraneous Data - Consistency, Uniqueness):

278 Extraneous data value(s) at index(es): [(0, '1'), (1, '1'), (2, '0'), (3, '1'), (4, '0'), (5, '1'), (6, '1'), (7, '1'), (8, '1'), (9, '1'), ('...', '...'), (268, '1'), (269, '1'), (270, '1'), (271, '1'), (272, '1'), (273, '1'), (274, '1'), (275, '1'), (276, '1'), (277, '1')] (displaying only the first and last 10 items)

Frequency Distribution:

student Frequency

1 216

0 62

Range of Values: (0 to 1)

hasPosition (position):

Categorical format: All 278 values are correctly categorical.

Categorical format with 5 unique value(s):

Category Frequency

0 226

Faculty 40

Faculty\_adj 6

Faculty\_aff 3

Faculty\_eme 3

inPhase (varchar):

String format: All 278 string values are valid.

String range (lexicographical): (0 : Pre\_Quals)

yearsInProgram (years):

Numerical >=0 format: Error(s) found:

DQI #17 (Wrong Data Type - Consistency):

140 Non-numeric value(s) at index(es): [(3, 'Year\_2'), (5, 'Year\_5'), (6, 'Year\_7'), (7, 'Year\_10'), (8, 'Year\_3'), (9, 'Year\_3'), (10, 'Year\_1'), (11, 'Year\_1'), (12, 'Year\_5'), (15, 'Year\_1'), ('...', '...'), (265, 'Year\_1'), (266, 'Year\_3'), (267, 'Year\_7'), (269, 'Year\_3'), (271, 'Year\_5'), (272, 'Year\_4'), (274, 'Year\_5'), (275, 'Year\_2'), (276, 'Year\_5'), (277, 'Year\_4')] (displaying only the first and last 10 items)

Range of values: (0.0:0.0).

YDATA Profiling results: It just found out that the PK column 'p\_id' has unique values:

