Adventures in Data Profiling

Jim Harris
Blogger-in-Chief
www.ocdqblog.com

Jim Harris

Jim Harris

Blogger-in-Chief

www.ocdqblog.com

E-mail
jim.harris@ocdqblog.com

Twitter
twitter.com/ocdqblog

LinkedIn
linkedin.com/in/jimharris

Let the Adventures Begin . . .

This will be a vendor-neutral presentation:

- Focusing on general methodology of data profiling and common functionality of data profiling tools
- Discussing how a data profiling tool helps automate some of the work needed for preliminary data analysis
- Reviewing fictional data and results produced by a fictional data profiling tool to illustrate basic concepts

Understanding Your Data

- Understanding your data is essential to using it effectively and improving its quality
- You need a reality check for the perceptions and assumptions you have about the quality of your data
- You need to prepare meaningful questions to ask your business analysts and subject matter experts
- There is simply no substitute for data analysis

Profiling Your Data

Data profiling includes many types of analysis such as:

- Verify data matches the metadata that describes it
- Identify representations for the absence of data
- Identify potential default and invalid values
- Check data formats for inconsistencies
- Assess domain, structural, and relational integrity

Getting Your Data Freq On

- Data profiling tools can help you by automating some of the grunt work needed to begin your data analysis

- One of their basic features is the ability to generate statistical summaries and frequency distributions for the unique values and formats found within your fields

Therefore, I like to refer to using a data profiling tool as:

"Getting Your Data Freq On"

Let Me Count The Ways

Field Name	NULL	Missing	Actual	Completeness	Cardinality	Uniqueness	Distinctness
Customer ID	0	0	3,338,190	100.00%	3,338,190	100.00%	100.00%
Account Number	0	0	3,338,190	100.00%	3,254,735	97.50%	97.50%
Customer Name 1	50,072	16,690	3,271,428	98.00%	2,997,864	89.81%	91.64%
Customer Name 2	2,450,670	53,077	834,443	25.00%	798,531	23.92%	95.70%
Tax ID	886,703	41,444	2,410,043	72.20%	2,120,837	63.53%	88.00%
Gender Code	1,204,060	50,264	2,083,866	62.43%	8	0.00%	0.00%
Birth Date	627,019	0	2,711,171	81.22%	25,275	0.76%	0.93%
Telephone Number	515,781	0	2,822,409	84.55%	2,624,840	78.63%	93.00%
E-mail Address	1,204,608	0	2,133,582	63.91%	2,037,570	61.04%	95.50%

NULL – record count of NULL values

Missing – record count of Missing values (i.e., non-NULL absence of data)

Actual – record count of Actual values (i.e., non-NULL and non-Missing)

Completeness – percentage calculated as Actual divided by total record count

Cardinality – count of the number of distinct actual values

Uniqueness – percentage calculated as Cardinality divided by total record count

Distinctness – percentage calculated as Cardinality divided by Actual

You Uniquely Complete Me

Input Metadata	
Field Name	Customer ID
Field Data Type	INTEGER
Field Length	10
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	3,338,190
Completeness	100.00%
Cardinality	3,338,190
Uniqueness	100.00%
Distinctness	100.00%
Data Profiling Additional Statistics	
Field Data Types	3
Field Length (MIN)	1
Field Length (MAX)	7
Field Value (MIN)	1
Field Value (MAX)	3338190
Field Formats	7

- Completeness and Uniqueness are useful in evaluating potential key fields and especially a single primary key, which should be both:
  - 100% Complete
  - 100% Unique

It's a Distinct Possibility

Input Metadata	
Field Name	Tax ID
Field Data Type	VARCHAR
Field Length	15
Data Profiling Summary Statistics	
NULL	886,703
Missing	41,444
Actual	2,410,043
Completeness	72.20%
Cardinality	2,120,837
Uniqueness	63.53%
Distinctness	88.00%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	9
Field Length (MAX)	9
Field Value (MIN)	000000000
Field Value (MAX)	999999999
Field Formats	1

Tax ID (Top 10 Field Values)	Count	Percentage
999999999	105,307	3.15%
000000000	56,545	1.69%
111111111	39,025	1.17%
888888888	28,557	0.86%
555555555	24,835	0.74%
824043548	7	0.00%
913010645	6	0.00%
786013577	5	0.00%
779044262	3	0.00%
795015738	3	0.00%

- Distinctness can be useful in evaluating the *potential* for duplicate records
- < 100% Distinct means some distinct actual values occur on more than one record

Gimme the lo down, Drill-down

Account Number	Tax ID	Customer Name 1	Customer Name 2
ER-852205406	824043548	Stephen Dedalus	NULL
ER-852205406	824043548	James Joyce	NULL
GF-574879711	824043548	Humphrey Chimpden Earwicker	NULL
GF-574879711	824043548	Richard Rowan	NULL
JJ-692571582	824043548	Polly Mooney	Bob Doran
JJ-333142972	824043548	Gabriel Conroy	NULL
JJ-853859650	824043548	Leopold Bloom	NULL
DQ-385867911	913010645	Soylent Consulting	Green, Incorporated
DQ-385867911	913010645	Green, Incorporated	Tab Fielding
DQ-385867911	913010645	Soylent Consulting	William R. Simonson
FB-578343697	913010645	Robert Thorn	NULL
FB-578343697	913010645	Sol Roth	NULL
GH-802359541	913010645	Sol Roth	NULL
BS-515199781	786013577	Peter and Lois Griffin	NULL
BS-515199781	786013577	Peter Griffin	NULL
NB-511232925	786013577	Stewie Griffin	Brian Griffin
NB-511232925	786013577	Lois Griffin	Stewie Griffin
FG-628337043	786013577	Meg Griffin	NULL
OC-238075019	779044262	Mr. and Mrs. Robert Frost	NULL
OC-238075019	779044262	Robert Frost	NULL
OC-238075019	779044262	Elinor Frost	Robert Frost

Freq'ing Distribution of Values

Field Name	NULL	Missing	Actual	Completeness	Cardinality	Uniqueness	Distinctness
Postal Address Line 1	196,536	5,193	3,136,461	93.96%	2,886,753	86.48%	92.04%
Postal Address Line 2	2,349,569	42,966	945,655	28.33%	875,578	26.23%	92.59%
City Name	171,517	15,171	3,151,502	94.41%	29,876	0.89%	0.95%
State Abbreviation	723,865	0	2,614,325	78.32%	72	0.00%	0.00%
Zip Code	925,591	0	2,412,599	72.27%	48,731	1.46%	2.02%
Country Code	0	0	3,338,190	100.00%	5	0.00%	0.00%

Country Code (Field Values)	Count	Percentage
US	1,966,983	58.92%
United States of America	741,776	22.22%
United States	555,773	16.65%
CA	71,867	2.15%
USA	1,791	0.05%

- Frequency distribution of values is useful for fields with a low cardinality
- Extremely low cardinality *might be* an indication of default values

Reviewing the Top *N* List

Input Metadata	
Field Name	Birth Date
Field Data Type	DATE
Field Length	10
Data Profiling Summary Statistics	
NULL	627,019
Missing	0
Actual	2,711,171
Completeness	81.22%
Cardinality	25,275
Uniqueness	0.76%
Distinctness	0.93%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	10
Field Length (MAX)	10
Field Value (MIN)	02-28-1929
Field Value (MAX)	12-21-2012
Field Formats	1

Birth Date (Top 10 Field Values)	Count	Percentage
04-16-1953	359	0.01%
05-25-1959	284	0.01%
12-19-1955	225	0.01%
07-21-1980	189	0.01%
07-04-1971	185	0.01%
02-28-1929	180	0.01%
11-25-1945	177	0.01%
07-13-1939	165	0.00%
10-16-1934	162	0.00%
01-22-1987	159	0.00%
Birth Date (Top 10 CCYY Values)	**Count**	**Percentage**
1955	112,575	3.37%
1964	103,230	3.09%
1953	101,457	3.04%
1966	86,615	2.59%
1971	85,704	2.57%
1959	82,804	2.48%
1975	58,542	1.75%
1929	54,983	1.65%
2011	48,424	1.45%
2012	44,051	1.32%

Reviewing the Top *N* most frequently occurring values

Freq'ing Distribution of Formats

Customer Name 1 (Top 20 Field Formats)	Count	Percentage	Sample Field Value
Given-Name Family-Name	442,892	13.27%	Peter Griffin
Given-Name Given-Name Family-Name	389,534	11.67%	Homer Jay Simpson
Given-Name Given-Name Given-Name	324,013	9.71%	Harris Edward James
Given-Name Given-Name	182,817	5.48%	Joel Ethan
Given-Name Single-Alpha. Family-Name	161,020	4.82%	Theodore D. Kerabatsos
Family-Name Given-Name	146,876	4.40%	Marlowe Christopher
Given-Name Single-Alpha Given-Name	135,833	4.07%	Daragh O Brien
Given-Name Family-Name Family-Name	123,441	3.70%	Henry Walton Jones
Family-Name, Given-Name	112,815	3.38%	Shakespeare, William
Given-Name Family-Name Business-Type	93,953	2.81%	Alan Smithee LLC
Alpha Business-Word	91,986	2.76%	Soylent Consulting
Given-Name Single-Alpha. Business-Word	88,263	2.64%	Joe T. Plumber
Alpha, Business-Type	81,362	2.44%	Inconceivable, Incorporated
Given-Name Family-Name—Family-Name	82,302	2.47%	Lorraine Baines-McFly
Family-Name Business-Word Business-Type	75,045	2.25%	Cooper Construction Company
Single-Alpha.Single-Alpha. Family-Name	67,911	2.03%	J.D. Salinger

Frequency distribution of formats is useful for fields having both a high cardinality and free-form values

Unlocking the Combination

Input Metadata	
Field Name	E-mail Address
Field Data Type	VARCHAR
Field Length	100
Data Profiling Summary Statistics	
NULL	1,204,608
Missing	0
Actual	2,133,582
Completeness	63.91%
Cardinality	2,037,570
Uniqueness	61.04%
Distinctness	95.50%

Combination of values and formats can help with unlocking the mystery of more complex fields

E-mail Address (Top 10 Field Formats)	Count	Percentage
USER@DOMAIN.TLD	1,173,470	35.15%
USER.USER@DOMAIN.TLD	525,408	15.74%
DOMAIN.TLD	170,686	5.11%
USER@DOMAIN.DOMAIN.TLD	93,811	2.81%
USER@DOMAIN@TLD	42,671	1.28%
USER@DOMAIN	32,003	0.96%
USER.USER@DOMAIN.DOMAIN.TLD	12,521	0.38%
USER-USER@DOMAIN.TLD	2,178	0.07%
@DOMAIN.TLD	1,161	0.03%
USER@DOMAIN-DOMAIN.TLD	326	0.01%
E-mail Address (Top 5 DOMAIN Values)	**Count**	**Percentage**
gmail	346,753	10.39%
yahoo	277,402	8.31%
hotmail	221,921	6.65%
aol	177,537	5.32%
ocdqblog	1,535	0.05%
E-mail Address (Top 5 TLD Values)	**Count**	**Percentage**
com	1,653,061	49.52%
org	246,752	7.39%
net	80,228	2.40%
ca	35,939	1.08%
uk	15,375	0.46%

. . . the Adventures Conclude

What can just your analysis of data tell you about it?

- Understand your data better by first looking at it from a starting point of ***blissful ignorance and curiosity***
- A tool can help automate some of the grunt work, but the ***actual data analysis*** can not be automated
- Your analytical goal is *not* to try to find answers, but to ***discover the right questions***