

**Information
Builders**

WebFOCUS





Data Profile Analysis
For

CRM data

8th October, 2009

Mr. John Smith
Marketing Manager

Summary of findings

This represents the output of a three hours spent profiling the CRM data as kindly provided by John Smith from XYZ Imports UK Ltd. This was supplied as a 20Mb csv file containing 11,203 rows of data.

This report is not intended as a complete profiling report. Below is a summary of some key findings.

Key Findings:

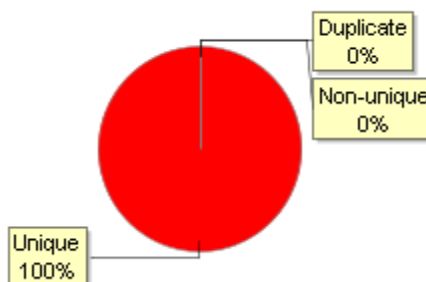
- The 11,204 records all have a unique identifier. Only 9 duplicates were found in the sample data.
- 47% of the field containing gender were populated with the value "null"
- For the field Title Group, 12% of the data contained the value "unknown"
- For the Field "ContEmail", 4% of the records contained no email address; 2% of the records were duplicates

Profiling Results

1. Field "ContactID"

"

Each record should have a unique ContactID field, and the CRM data quality is good in this regard. Of the total 11,204 records, only 9 contain duplicate fields.



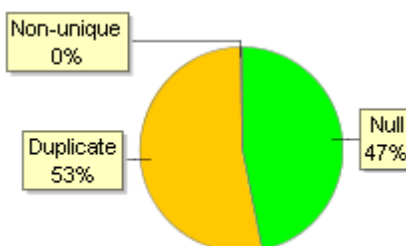
It seems there are 9 ContactID IDs for which a pair of records exist. Using iWay DQC drill-down functionality it was possible to view each of these pairs.

All other fields are unique. There are no missing values (Nulls).

A concern is that while the vast majority of ContactID fields contain text of 31 or 32 characters, there are some records with values less than this. This suggests that some of the records are incorrectly formatted, not adhering to the correct structure for a ContractID field.

2. Field "Gender"

Analysis of the Gender column shows that 47% of the values are null, implying that for 47% of records, gender is unknown. Of the 53% of records for whom gender is specified, 42% are male, and 10% female.



We are not aware of a resultant cost to the business related to this high volume of unknowns. However, with iWay DQC it will be possible to reduce this figure down to zero (or near zero). iWay DQC may be used to determine gender for each record based on data held in other columns. For example, the **Salutation** column containing titles has a populated rate of 75%. Many of these titles (Mr, Mrs, Ms etc) may be used to determine gender. The **FirstName** column (populated rate of 100%) may also be used to determine gender; with iWay DQC we can create a lookup table containing common first names along with associated gender.

These methods may be used to ensure that gender is known for all (or nearly all) records.

3. Field “Title Group”

This column indicates which executive tier a contact belongs to.

Data Quality for this column is good. All fields are populated.

However, the high percentage of ‘Unknown’ is a concern.

The image above shows the breakdown of contacts, nearly 40% of contacts in CRM are in the ‘Mid-Management’ category.

Value	Count	Percentage
Mid-Management	4,295	38.33%
Corporate Executive	3,834	34.22%
Non Management	1,765	15.75%
Unknown	1,310	11.69%

4. Field “Title “

5% of fields in this column are null, indicating that for 1/20th of the records in CRM, there is no indication of position within the company.

These contacts are therefore of very little value in a sales and marketing context.

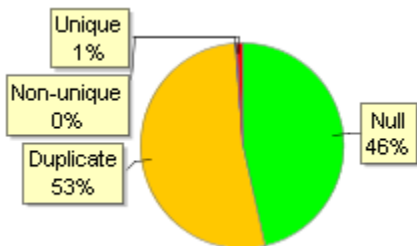
There is also an evident lack of naming standards used in these data; for example we see a percentage of “V.P. Sales” as well as “Vice President – Sales”.

IT Manager	330	2.95%
Financial Director	232	2.07%
Non Management	206	1.84%
Finance Director	177	1.58%
IT Director	173	1.54%
Middle Management	142	1.27%
Managing Director	131	1.17%
Marketing Manager	128	1.14%
Financial Controller	116	1.04%
Marketing Director	110	0.98%
Head of IT	87	0.78%
HR Manager	85	0.76%
HR Director	84	0.75%
Operations Director	73	0.65%
Corporate Executive	63	0.56%
Directorate Manager	63	0.56%

We recommend that steps be taken to remedy this. It may be possible to configure rules within iWay DQC to determine titles for these contacts based on other factors (such as location, department etc)

5. Field “Department”

For nearly half of all CRM contacts, their department is unknown. This renders the CRM data of little value for targeted marketing activities; should you wish to filter outreach programmes to target only those in HR departments for example, the CRM data would not allow you to do this efficiently.



Another problem with these data is the lack of standards for department names. Of the records for whom department *is* known, “Information Technology” is, by far, the most prevalent entry (at 22%). However, there is another separate group of records listed as “IT”, which should obviously come under the same classification. Similarly, there are other duplicate groups, such as “

We recommend that steps be taken to remedy this. iWay DQC may be used to enrich existing data by establishing rules in DQC to determine the department for each contact based on other factors.

Of the records for whom department *is* known, “Information Technology” is, by far, the most prevalent department, followed by Finance.

6. Field “Company Name”

Data Quality is good for this column. There are no missing values.

Not-nulls count:	11,204	100%
Nulls count:	0	0%

Of the 11,204 records in CRM, we see that these relate to 2,022 separate companies (as indicated by the ‘distinct values’ field).

Distinct values:	2,022	18%
Duplicate values:	9,182	82%
Unique values:	1,014	9%
Non-unique values:	1,008	9%

7. Field “Company ID”

For this column, we would expect that of the total 11,204 records, we should find 2,022 different Company IDs (since there were that many distinct company names). However, we see that there are 2023 distinct company IDs listed.

Not-nulls count:	11,204	100%
Nulls count:	0	0%

Distinct values:	2,023	18%
Duplicate values:	9,181	82%

Unique values:	1,015	9%
Non-unique values:	1,008	9%

This indicates that for at least one company listed in the CRM data, there is more than one ID. This indicates that the Company ID field is not reliable. For example, if you were to extract all data for a particular company using Company ID as the filter, you may not retrieve all data linked to that company, since there may be more than one company ID used for that company.

8. Field “Contact Phone”

The image to the left displays the iWay DQC analysis report on *format* of the data held within the Contact Phone column.

The results here are alarming.

Nearly a quarter of all contacts have no associated phone number.

Of more concern however, is the fact that, the fields which are populated seem to contain unreliable data. A phone number should be of uniform format, or structure, containing a certain number of digits. The results displayed here

Value	Count	Percentage
NULL	2,595	23.16%
#	7,357	65.66%
# # #	955	8.52%
# #	172	1.54%
D	30	0.27%
#.#	14	0.12%
#/#	14	0.12%
#.#.#	11	0.10%
#-#-#	8	0.07%
# D # #	6	0.05%
# # D	4	0.04%
# #-#	4	0.04%
# # # #	3	0.03%
(#) #-#	2	0.02%
(#)#-#	2	0.02%
[#]#	2	0.02%

indicate that the format stored in this column is of no fixed format. Very few of these fields contain valid phone numbers.

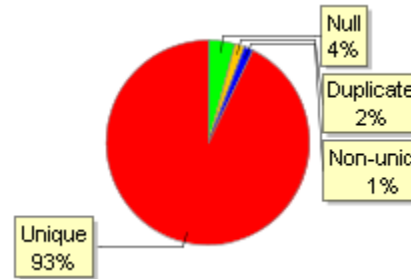
9. Field “ContEmail”

The CRM email data contains several serious data quality issues.

There is no email content for 4% of contacts.

An alarming 2% of email entries are duplicates. This indicates a serious problem. A CRM record refers to a person within an

organization. Each email address stored within CRM should therefore be unique (as duplicate email addresses do not exist).



Using drill-down functionality, we can see email addresses for which many contacts have been assigned. Unfortunately, there are many more email addresses shared by multiple contacts within the CRM data.

iWay DQC could be used to remedy this by passing all records through a data validation routine which would verify the email address against the first name and last name stored against that contact. If the email address does not correspond to the names, iWay DQC could cleanse the data by constructing a valid email address for that contact.

Another issue with email addresses is highlighted by the format analysis on the email address data. This indicates that a very high number of email addresses do not contain a ‘@’ character. This obviously means the email addresses are invalid.

Again, this could be remedied using iWay DQC and automated email address generation routines.

10. Field “Address Data”

The **City** column is well populated, showing that London is by far the most prevalent city location (2,037 records have ‘London’). The next most common city is Dublin with 281 records. This is surprising, given the companies focus on the UK market. It seems there are many untapped contacts in the Republic of Ireland.

The author of this report wonders if we should be engaging our talented presales members on these accounts.

Other address data is poor. The **PostalCode** column for example contains no valid data. 97% is null, and of the remaining 3%, none of the content refers to valid postcodes.

This indicates that the address data held within CRM is not fit for purpose.

iWay DQC may be employed here to improve the quality of address data within CRM, for example, by utilizing lookup services to enrich the existing partial address data, or by matching addresses across contacts who belong to the same company/branch.

Technology used

iWay Data Quality Center version 5.2.3

On a Dell 630 Latitude laptop running Windows XP, 2.6GHZ, 4GB RAM

Profiling time was approx 20sec for the complete set of CRM data.

A representative sample of the output of the iWay DQC is included in the screen shot below.

