**IMPUTATION METHODOLOGY**

Imputation groups

A combinaison of ISIC and SIZE

*ISIC*

18 groups using the ISIC labels

*SIZE*

Three size groups (based on total revenue – BR8) will be created within each of the 18 industry groups. The 33e and 66e percentile of BR8 will be calculated to determine the boundaries of each size group.

Rules:

There should be at least 30 donors in each ISIC groupings (10 donors in each ISIC x SIZE groups). If not, SIZE is not used.

Specifications (example for the Business Expenses section)

**TABLE ImputationBE**

|  |  |  |
| --- | --- | --- |
| **EDITS** | 1 | 2 |
| BE15 is greater than 0 | Y | N |
| **ACTIONS** | | |
| Identify the nearest neighbour as donor | 1 | 1 |
| BE15 (recipient) = BR8 (recipient) \* BE15 (donor) ⁄ BR8 (donor) |  | 2 |
| BE1 (recipient) = BE1 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 2 | 3 |
| BE2 (recipient) = BE2 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 3 | 4 |
| BE3 (recipient) = BE3 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 4 | 5 |
| BE4 (recipient) = BE4 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 5 | 6 |
| BE5 (recipient) = BE5 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 6 | 7 |
| BE6 (recipient) = BE6 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 7 | 8 |
| BE7 (recipient) = BE7 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 8 | 9 |
| BE8 (recipient) = BE8 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 9 | 10 |
| BE9 (recipient) = BE9 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 10 | 11 |
| BE10 (recipient) = BE10 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 11 | 12 |
| BE11 (recipient) = BE11 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 12 | 13 |
| BE12 (recipient) = BE12 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 13 | 14 |
| BE13 (recipient) = BE13 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 14 | 15 |
| BE14 (recipient) = BE14 (donor) \* BE15 (recipient) ⁄ BE15 (donor) | 15 | 16 |

Algorithm to find the nearest neighbour

\*This step is not part of the imputation process. It will be run after the Preliminary E&I process once the usable records are identified.

1) Create imputation groups\*

* 1. For each of the 18 industry groups, compute the 33e and 66e percentile to determine the three size groupings.
  2. If there are less than 30 donors in the industry group, create only one size group (size = 0)
  3. If there are at least 30 donors, create three size groups using the percentiles (size = 1, 2 or 3)
  4. For each record, create a derived variable called ISICSizeGroup which is the concatenation of the industry group and the size group.

2) Identify the nearest neighbour

\* This step is module-specific. This means that this process has to be run for each module.

2.1 Identify the set of donors using the imputation flags derived in decision tables for that

module (DV\_Impute\_module ≠ 1).

2.2 Within each imputation group, identify the nearest neighbour by comparing the recipient

revenue against each donor revenue using an iterative process. Amongst those, identify the

donor with the minimum difference where DV\_. Mathematically speaking, this translates in:

Nearest neighbour of a recipient in an imputation group is the donor that minimizes

Min |BR8(recipient) – BR8(donor)| where donor ∈ imputation group

Summary (to be completed)

|  |  |
| --- | --- |
| **Imputation Flags** | **Description** |
| DV\_Impute\_BR\_Details = 1 | Imputation of BR1-BR7 |
| … |  |
| … |  |
| DV\_Impute\_WholesaleRetail\_Details\_Sales = 1 | Imputation of GPRW1S-GPRW15S |
| DV\_Impute\_WholesaleRetail\_Details\_Cost = 1 | Imputation of GPRW1C-GPRW15C |
| … |  |