

1. Data for the previous year is submitted to the Statistical Office by the end of the first quarter of the current year. The data is usually presented in excel format in several excel sheets (*Births Illegitimate Kingstown, Births Illegitimate Out District, Births Legitimate Kingstown and Births Legitimate Out district*). The Late registrations are clearly marked by a comment '*Late Registration*'. All data are placed in one excel sheet under the following variable names:
 - a. Stats_ID - (this is left blank until the data is cleaned and ready for SPSS)
 - b. PlaceOfBirth - this is a string variable: options are 'Kingstown' or 'out district'
 - c. Legit_Illegitimate - this is a dichotomous variable: 1 = illegitimate & 2=legitimate
 - d. Entry_Number - this is the number used by the treasury as a unique identifier
 - e. Gender - this is a string variable: options are 'M' for males and 'F' for females
 - f. Mother_Age - this is continuous (single years)
 - g. Mother_Age_Group - this categorical variable is generated from 'Mother_Age'
 - h. Parity - Order of birth
 - i. Occupation
 - j. Isco_Code - ISCO Rev 08 is used to code occupation into broad categories (single digit). Code '11' is applied to those who did not provide a response to this question. Code '12' is used for those who indicate 'unemployed' or 'student', note however that a 'student nurse' is treated differently. Persons in these categories have no occupation since they are not employed.
 - k. Address
 - l. CensusDivision - a census division code of '01' to '13' is applied based on address. Code '14' is used if address is unknown OR an overseas address. In reporting all unknown and overseas addresses are merged as 'unknown'.
 - m. Mother_First_Name
 - n. Mother_Middle_Name
 - o. Mother_Last_Name
 - p. DOB -date of birth of the child
 - q. Multiple_Birth - If twins a code of 2 is used, if single birth 1, etc.
 - r. Month - codes 1 to 12 are used for months January to December
 - s. Year
2. The mother's full name, 'Entry_Number', and 'DOB' of child are used to identify duplicates. All duplicates are then removed from the data file. Sometimes it is necessary to look at other variables such as 'sex' and 'placeOfBirth' to determine which cases are duplicates. In most cases, duplicates will have one '*Out-District*' code and one '*Kingstown*' code. In a few cases, duplicates can be seen where there are late registration. That is, the birth is noted in Kingstown (place of birth) then the parent registers in their own district - both cases may end up in the datafile.

CAUTION - multiple births are not seen as duplicates and must not be discarded from the database. Multiple births will have same name of Mother, Same DOB, but different entry numbers. Sometimes there may be a duplicate in a set of twins - giving the impression there were 4 live births.
3. After all duplicates are removed from the database, the result is checked against the count of live births (twins, triplets, single births) provided by the MOH.
4. If it is determined that the data is of sound quality, the data base is added to the other Live birth data in Excel and SPSS for extraction of tables. The Staff has an option as to what software is used to manipulate tables for extraction. SPSS is quicker and allows for further analysis.

HAPPY DATA CLEANING