Type of errors in this dataset and the way to clean

1. Multiple representation.

   For example, Department of Agriculture, DoA and Agriculture department are the different representations for the same thing, but computers treat them as different words.

   They way to solve it is that find the words represent same thing and then merge them into a same name. I used Openrefine to do this.

2. Data format.

   This is mainly for some special data columns. For example, 02/03/2013 and 02-03-2013, which may affect the importing of data and the later analysis.

   We can use openrefine to detect different format and fix them.

3. Different ranges for numeric data.

   The real value of data may different because of the unit. For instance, 1 Kg and 1g or 100$ and 100$m. The data may be collected by some human error and different departments may use different units to record.

   To detect this error, we need to use some tools. In openrefine, we can use some mathematical tool, such as view them on a log scale, and then we can check if there is any outlier. But next we need to judge if it is a mixed use of numeric scales by ourselves.

4. Redundant records.

   Some records in the dataset have same meanings with other records and it may have some bad influence for the next processing. In this dataset, there are some summation records that compute the total data in a subset, which is useless because we already have the whole dataset. In addition, the record like that may have different structure and it will affect reading data.

   The way to fix it is remove the redundant records. It does not cause the data missing and but make the dataset cleaner. I did this by Openrefine.

5. Spelling errors.

   That may be caused by human mistake. It is good to use Openrefine to do cluster for the whole column. For example, we can find that Department of Agraculture and Department of Agricultrue and then merge them into the correct one.

6. Records structure.

   There are some records that do not in the same position with other records. For example, there are some blanks in the front and each cell is misaligned with that in other rows. That could be adjusted by some data operating software like Excel.