Data CLeaning Report

Keen Koalas – Nautical Narwhals

INFM600 - Section 0101

Saba Aldughaither

Mayuresh Amdekar

Eris Mei

Himanshu Sawant

**Data Source:**

The Pesticides Data Program (PDP) is a national pesticide residue monitoring program, which is implemented by United States Department of Agriculture, and it produces the most comprehensive pesticide residue database in the U.S. This dataset contains PDP sampling and residue testing data and results along with file layouts and reference lists that can be used to interpret the standardized codes used in the PDP dataset. The dataset contains the data for the years ranging from 1992-2014. The Agricultural marketing service has tested for pesticide residue on over ten thousand samples of agricultural goods sold in the U.S every year. These datasets include details such as where and when the item samples were collected, the types of processing the products had undergone, the types and amounts of residue, and whether the products had organic or pesticide-free claims. The monitoring program division administers PDP activities, including the sampling, testing, and reporting of pesticide residues on agricultural commodities in the U.S. food supply, with an emphasis on those commodities highly consumed by infants and children. The program is implemented through cooperation with state agriculture departments and other federal agencies.

Dataset Citation:

United States Department of Agriculture, Agricultural Marketing Service. (2016). PDP Databases, 1992-2014. [ZIP archives with data files]. Retrieved September 26, 2016, from: <https://www.ams.usda.gov/datasets/pdp/pdpdata>

**Licensing:**

The Agricultural Marketing Service (AMS) of the USDA specifies on its privacy page that all information on its site is public and can be copied or distributed if credit is given.

United States Department of Agriculture, Agricultural Marketing Service. (2016). Privacy Statement. Retrieved November 2, 2016, from: <https://www.ams.usda.gov/about-ams/privacy>

**Metadata:**

This dataset consists of the details of the food samples which were tested from 1992-2014. There are standardized codes in the dataset which are explained in the metadata. The metadata consists of:

1. An “Instructions” file which explains how to load the dataset in Microsoft Excel or Microsoft Access and also a description of all the files which are included in the data for a particular year such as the samples file, results file, data dictionary, etc.

2. The “Data Dictionary” file for the samples and results for each year which describes the column headers, the data type of the column and the number of samples and results for each year.

2. The “Reference Tables” file, which can be used to interpret the standardized codes in the dataset (for example, there are reference tables describing the codes for commodity names, commodity type, country name, origin of sample, pesticide name, etc.).

**Rationale:**

The dataset contains the annual PDP sampling and residue testing data from 1992 to 2014. However, for a single year, say 2014, the sampling data contains 10,619 rows and the results data contains 2,210,744 rows. To limit the scope of the project, we have decided to work with the samples and results for only one commodity type: Apples (often ranked #1 on the Dirty Dozen list for pesticide residue), and for only the latest year: 2014. However, to answer one research question related to the trend of concentrations, we are using the data from 2004 along with that of 2014 to plot a trend over a 10-year interval. Also, we have decided to drop the columns with mostly blank values and are not used in any of the analysis questions. We also combined two columns with country related information to aggregate their data into one column to create a leaner database. Lastly, the missing values in a few rows were filled by comparing the data for the same attribute in corresponding data from a different year.

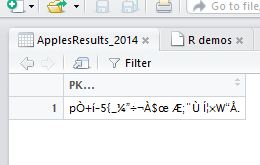
**Issues Encountered with the Data:**

1. Extraction of data into format usable in R:

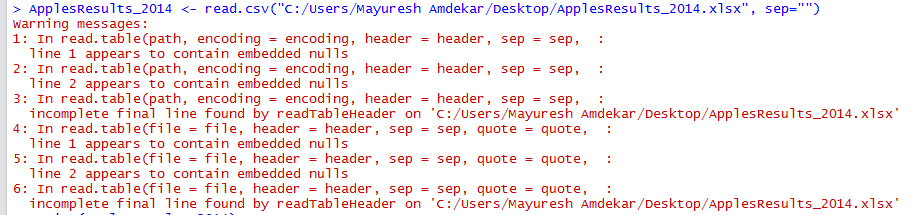
The dataset is available in the form of a MS Access database with accompanying data in “.txt” files. With MS Access, this data can be extracted into only the formats “.xls\*”. However, to use the dataset in R, the dataset needs to be in CSV (Comma Separated Value) format. Therefore, we converted the data directly from the “.txt” files that accompanied the MS Access database into CSV as outlined in the data cleaning steps below.

The following is the error we encountered when we tried to import an “.xlsx” file as dataset into R:

Data Snapshot:



Console:



1. Blank/NULL values

There are a few columns in the dataset which have considerable numbers of blank/NULL values. The following are the columns for “Samples” and “Results” that have mostly blank values:

* Samples Data:
  + The attribute “SOURCE\_ID” is blank for 158 rows / 177 rows of the Samples Data for 2014. According to the metadata, it is used to make a sample unique. However, all the samples have a unique value in the SAMPLE\_PK attribute, which is the primary key for the database. Therefore, this attribute can be dropped from the table.
  + The attribute “COUNTRY” will contain a value only when the “ORIGIN” attribute has a value of '2', i.e. the variety is imported from another country. If ORIGIN has a value of '2', the COUNTRY attribute will have the country code of the country of origin, e.g. '660' for New Zealand. As most of the samples originated in the United States, the “COUNTRY” column has many blank/NULL values. For the 2014 data, 176 rows /177 rows have blank values for the Country attribute. We can replace the value in the “ORIGIN” attribute with the name of country (e.g. “United States” for '1', “New Zealand” for '2' if its corresponding COUNTRY value is '660') and drop the “COUNTRY” column.
  + The fields GROWST, PACKST, and DISTST have large numbers of blank values. As these fields are not used in the answering of research questions, they can be safely dropped.
* Results Data:

The attributes CONFMETHOD2, ANNOTATE, and QUANTITATE have blank values for all 38,792 rows of apple data from the year 2014. As these attributes are not used in analysis, they can be dropped from the dataset.

**There are similar observations for the “Samples” & “Results” data for 2004.**

1. Missing Values

A Considerable number of rows (106/744) have a null value for the “VARIETY” attribute for the Samples data of 2004. The data for the year 2014 contains a few rows with the value “Unknown” for the attribute “VARIETY”. The same value can be used to fill the empty attributes of the data for the year 2004.

**Steps followed to clean the data:**

Data Cleaning is done in two phases:

1. Creation of CSV File for the data
   1. Open a new, blank file in Excel.
   2. Under the “Data” tab, click on the “From Text” button.
   3. Select the Samples.txt file in the window that pops up for text import, and click the “Import” button.
   4. Under Delimiters on the second page of the Text Import Wizard, check the box for “Other” and enter “|” as the delimiter, then click the “Finish” button.
   5. Add the appropriate headers for each column by referring to the “PDP DataDictionary” file that accompanied the dataset.
   6. Save the file as CSV with a **temporary name**.
   7. Filter the data for apples by selecting value for attribute “COMMOD (Commodity type)” as 'AP'.
   8. Copy the complete data after filter is applied and paste it into another blank CSV file.
   9. Save the new CSV file with an **appropriate name**.

The above steps are followed for creating CSV files for Samples and Results data for both 2004 and 2014.

1. Execute R Script to drop the unneeded columns mentioned in the “Blank/NULL values” section above.