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 9, 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 year 1992 to 2014. However, for a single year, say 2014, the sampling data contains 10619 rows and 18 columns and the results data contains 2210744 rows and 16 columns. For the year 2004, the sample data contains 13208 rows and 18 columns and the results data contains 1281438 rows and 16 columns. 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) 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 few rows were filled by comparing the data from the same attribute in corresponding data for a different year.

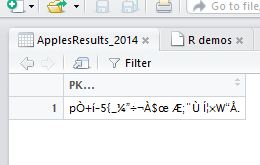
**Issues Encountered with the Data:**

1. Extraction of data into format usable in R: -

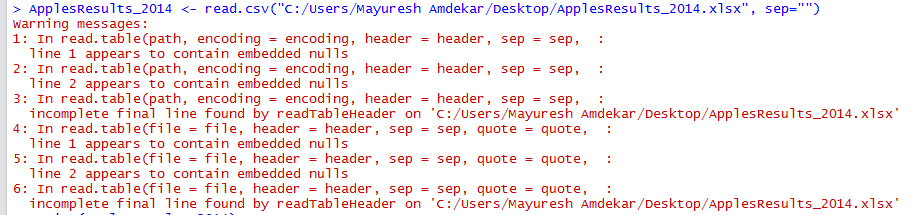
The dataset is available in the form of a MS Access database with accompanying data in “.txt” files. With MS Access, this data can be extracted into only the formats “.xls\*”. However, to use the dataset in R, the dataset needs to be in CSV (Comma Separated Value) format. Therefore, we converted the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

Data Snapshot: -



Console: -



1. Blank/NULL values

There are a few columns in the dataset which have considerable number 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. As per 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 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 have 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 Country attribute. We can replace the value in the “ORIGIN” attribute with the name of the country e.g. “United States” for '1’, “New Zealand” for '2' if its corresponding COUNTRY value is '660' and drop the "COUNTRY" column. Also in the samples data for 2004 there is a row where the value for Origin is 3, we replaced it with the value “Unknown” as given in the data dictionary.
  + The attributes GROWST, PACKST and DISTST which represent the state postal code for location of grower, packer and distributor respectively have large number of blank values. As these attributes are not used in the answering of research questions, they can be safely dropped.
* Results Data:

The attributes CONFMETHOD2, ANNOTATE and QUANTITATE which represent the code for secondary confirmation method, code for annotated information about positive residue and code for quantitative method have blank values for all the 38792 rows of apple data from year 2014. As these attributes are not used in analysis, they can be dropped from the data set.

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

1. Missing Values

A considerable number of rows (106/744) have 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.

1. Unification

In the 2014 Samples data, "Variety" is a free form text field, so there are various versions of variety names entered that had to be unified. A new column named "Grade" is manually added to store the fancy descriptive grade information in the unification process. (Eg. “Red Delicious Fancy” Variety is changed to Red Delicious and the Grade is written as “Fancy”). By doing this, apples of that variety can be grouped together for analysis regardless of the grade and even the apples can be grouped on the basis of the grade.

**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.

After data cleaning, the number of records which we obtained to carry out analysis for the pesticide data of 2004 and 2014 are as follows:

Samples 2004: 744 rows 14 columns

Results 2004: 96591rows 13 columns

Samples 2014: 177 rows 14 columns

Results 2014: 38792 rows 13 columns

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

Word Count: 1269