Data Cleaning Project – Phase I

Team Members:

Goutam Debnath - goutamd2@illinois.edu
Debabrata Biswas - dbiswas3@illinois.edu
Rohit Narula - rnarula2@illinois.edu

1. Develop target/main Use Case(U1).

One of the Use cases of could be:
Give me all the clusters (group of close by cities near me) where there is a farmers' market which provides *both egg and vegetables*?

The user will want to minimize his transportation time and at the same time get all the products he needs.

2. Minor Use Cases:

Data cleaning is not necessary

A use case which will require very less, or no cleaning or data preprocessing will be answering the following question:

1. How much of each payment methods such as credit, WIC and others in overall marketplaces are available?

2. Find all the markets for given zip code where I can use credit card?

Data cleaning is not sufficient

One use case in which the data will never be good enough is if someone want to export the prices of all the goods a market provides by web crawling their website or social media link. The reason for this is that most of the links are not valid or not present which makes it impossible to be able to access all of them successfully and scrape some information of their web sites.

3. Describe the dataset

Structure and content of the dataset and quality issues that are apparent from an initial inspection

The farmers market dataset contains 59 columns. Broadly classifying columns as below for better understanding:

FMID is identifier for each farmer's market and contains 7 digits number.

Market name which is free text column and allows special characters.

Media Sites: Next five columns (Website, Facebook, Twitter, YouTube, OtherMedia) are supposed to contain information that will uniquely identify the market in one of the world wide web media sites. Most of the times this information is an URL to their page but there are no restrictions or validation so

that sometimes the value in these columns is free text. All the five columns support blank or null values.

Market Address: Next five columns (street, city, country, state, zip) are supposed to provide the address that will uniquely identify the market on the geographic map. All of them are free text columns, even the ZIP code column contains some special characters like '-'. However, one quality issue is that some of the names are lower case, some capital case, other have whitespaces before or after the string. This will cause issues if the user wants to do direct string comparison. Also, a good way to avoid that and establish unification over these names of countries, cities and streets can be for example to give each country a unique identifier and map the row that should belong to that country to the identifier instead.

5 Season Date Time: The following eight columns containing the date and time for every season most likely represent the periods when the market is opened. The initial assessment clearly shows that these columns also are not consistent in terms of format. There are occurrences where the period is stated as month A to month B and at the same time other occurrences where period is more granular for example from date to another date.

The X and Y columns represent latitude and longitude, although the names of these columns are not meaningful enough and the only way for a user to understand that is to see the column values. Another not so meaningful name is ' location'. The understanding is that it will contain map/geographic specific information, however it looks like it is more of a description of the place where the market is held.

Market characteristic: The next 35 columns are Boolean columns containing Y-true and N-false for given characteristic of given market. An easily distinguishable data issue is that sometimes they have a third or fourth value option like '-' or empty string which most likely indicates these tuples do not have neither false or true values for these columns.

Update Time: The last column (updateTime) representing date and time when the record has been updated also does not follow consistent datetime format. There are records with only year, records with full date time and records where month is present with a word instead of a digit.

4. List obvious data quality problems

Dataset contains lots of null data. Mostly, it lacks media attributes such as website, Facebook links and others. Although some location data like zip, street, city are missing, it can be replaced by using X, Y data which are assumed to be longitude and latitude of the market.

Too much season date and time are missing. Almost 95% of Season2 to Season4 date and time data are missing, which seem to be hard to find use case for those data. Also, information about what kind food is available in the market are missing largely.

Season date are not in the same format. They mostly follow DD/MM/YYYY format, but some are in Month DD, YYYY format.

Attribute county contains both lowercase and uppercase.

5. Devise an initial plan:

1) Data Cleaning methods and process with Open Refine - Goutam Debnath

With OpenRefine, data will be clustered if they are in similar text, or reformatted to keep consistency of data. Firstly, all column data should be trimmed and collapsed if they have consecutive white spaces. Next, county, city, States names are inconsistent. Some of them are in uppercase while others are not. They need to be converted into same format by clustering function. Some of SeasonDate columns contain various formats which should be fixed with the use of regular expression.

2) Applying more suitable solutions - Debabrata Biswas

Albeit OpenRefine is a great tool for cleaning data, there are some limitations. For Farmers Market data, the tool cannot help filling some of missing zip code. And by using 'uszipcode' package in python along with latitude, and longitude variables (which are 'x', 'y'), zip code can be filled.

Based on closest match from given latitude, and longitude, when city name of searched zip code and city name of the data matches, the data zip code is assigned with matched zip code. In case there are no matches for city names, zip code is given according to the closest distance.

3) Develop Relational Database Schema - Rohit Narula

Create Schema by importing the cleaned data from earlier steps into SQLlite database
Integrity constraints check

4) Create a Workflow diagram for the actual data cleaning steps - Goutam Debnath / Debabrata Biswas / Rohit Narula

Use the YesWorkflow on-line editor to create the workflow graph for whole process.