**Overview and Initial Assessment**

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

The farmers market dataset contains 59 columns. The first column FMID is identifier for each farmer’s market and contains 7 digits number. The second column is the market name which is free text column and allows special characters. The 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.

The 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.

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 in 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.

The next 35 columns are Boolean columns containing Y-true and N-false for given characteristic of given market. A 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.

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.

Some of the major issues in this dataset are the inconsistency of the format of data values for most of the columns. This is caused because of all the columns are free text columns and probably the software which inserts the data in this data source does not have any validations. The consumer of this data cannot rely on any kind of consistency and operations like comparison will always be tricky unless the consumer does not have any logic to convert the current column format to one of the ISO standardization formats. Moreover, some of the values are meant to mean the same thing but are written a little differently missing a letter or having special characters or having different casing. The data does not violate First or Second Normalization database forms because the FMID is unique and there are no duplicates of this key having different values as attributes. However there is still a lot of redundancy in terms of the way different types of products are associated with a market.

**Use case of the dataset**

A use case which seems to be very accurate is if a customer wants an answer of the following question: “Give me all the clusters (group of close by cities near me) where there is a farmers market which provides both seafood and vegetables?” The user will want to minimize his transportation time and in the same time get all the products he needs.

**Derive from it some data cleaning goals that can achieve the desired fitness for use.**

As of now, the way to fetch the desired data is to filter all the records that have ‘Y’ for Seafood and Vegetables and are currently functioning now (Select the current season date and time and check if the date and time are current) and then estimate the distance between the customer’s current location and the market location based on latitude and longitude and Haversine formula. The query should account the time it will take the user to arrive in the desired location and select only the closest markets which will be open at least 30 minutes after ETA of the user.

In order the query to be optimal the date sparsity should not be like it is in its current condition. For instance, the latitude and longitude should be accurate and able to form a cluster based on users’ location and trying to minimize the radius of the cluster. There are Geo Spatial and proximity search algorithms which can be applied here. Also, another difficult to query data is the date and time the market is opened. Having the concepts of seasons, it is hard to determine the boundaries of each season so that the query will know which column to check for datetime to match the user’s desired datetime. Basically, the query should go over all the 8 season/date/time columns and look for this time interval. Furthermore, the values in date and time columns follow 2-3 different formats which will make it difficult to select and filter without loading all the data into memory, converting these column’s values into a consistent format and then apply the filtering.

**Are there use cases for which the dataset is already clean enough?**

An use case which will require very less or no cleaning or data preprocessing will be answering the following question: “Find all the markets for given zip code where I can use credit card?”

Both zip and Credit columns are fairly consistent and queries running against them can be fast without additional data manipulation.

**Others for which it well never be good enough?**

One use case in which the data will neve 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.