**Introduction and initial assessment of dataset**

The farmersmarkets.csv file contains data including markets’ name, type of products available, time of operation, market location and websites.

The first row of the csv file is the column name and the first column is the FMID of each market, followed by market name which ideally should only contain strings without special characters.Then it lists all the websites and facebook, twitter and other media sites available, however, these columns contain values which are not consistent. Since some of them only include user name without valid URLs.

The next 5 columns contains the location information for the specific market. The quality issue for city and states is that some values are in lowercase while others might be in uppercase. Another issue is that special charaters also presents in those columns. Needless to say there are some extra white spaces before and after these strings.

Next 8 columns include data for operation dates. The seasonal dates are not all in the same standrad date format which makes it hard for user to query or select data following certain rules. Some entries are in date/mm/year format while others enter the period in month to month format(eg. July to Novemeber, 06/25/2014 to 09/30/2014). The next 2 columns include information for X and Y corresponding to latitude and longtitude values.

The rest columns representing the availablity of type of food provided in the market. Each entry is marked either Y or N. But Null values are also present in those columns. The last column in this csv file is the updated time for the data set. Most rows contain data formated in mm/dd/yr h:m:s format, but some rows only have year.

The major issue of this data set is inconsistency. Most values are entered without following specific rules such as ISO standard formats. This could make query data harder than expected and the results of these queries will not be conclusive enough for the user.

**Use case can be achieved from this dataset**

Even though the dataset has lots of flaws, user can still get correct result for questions like: Which Farmer’s market sell Seafood and accepts credit card payment? People these days might only bring their credit card with them and if they don’t bother to visit the bank and get cash before go buy seafood, they might want to know where is the that could let them purchase seafood with credit card.

**Use case that can be achieved after cleaning this dataset**

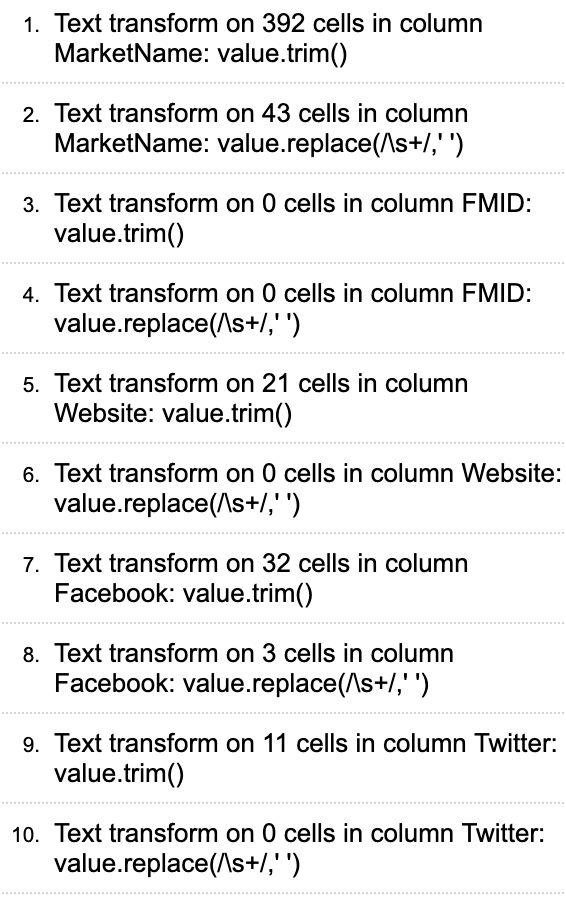
After cleaning the dataset, we hope to be able to answer questions like where is the closest market that I can go to buy flower and seafood with only credit card. By filtering data through its location and operation hour, we could get the closest opening market and then we can further select the data with flower and seafood entries to be ‘Y’ in order to answer the above question.

**Use case that can never be achieved from this dataset**

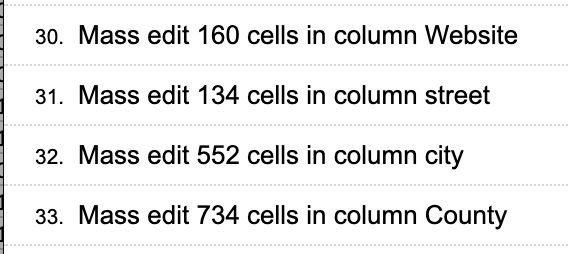
Among all the markets available, which market sells the cheapest flower? Since the dataset itself does not contain market prices for their products, we cannot get the price comparisions directly from this dataset.

**Data cleaning methods and processes**

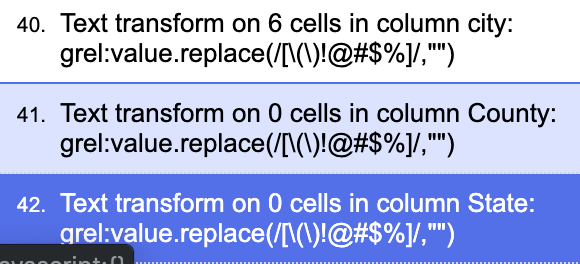
Trim extra white spaces from all the columns:



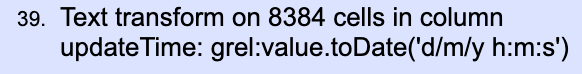
Cluster identical names together (website, city, state, country)



Remove special charaters (city, street)



Convert to ISO standard format (Update time)



Split into two columns(start date and end date for seasonal date 1)

