**CS513**: Theory and Practice of Data Cleaning – **Final Project (Summer 2021)**

Submission: Team37 NetIDs: sunilk2, vs24, amk8

# Abstract:

Table of Content

[Abstract: 1](#_Toc78399960)

[Introduction & Background 1](#_Toc78399961)

[Data Set 2](#_Toc78399962)

[Data Source 2](#_Toc78399963)

[Problem Statement 2](#_Toc78399964)

[Description of Dataset 2](#_Toc78399965)

[Data Exploration 4](#_Toc78399966)

[Data Quality Issues 4](#_Toc78399967)

[Use Case 6](#_Toc78399968)

[Other Potential Use Cases (Dataset "Clean Enough") 6](#_Toc78399969)

[Unrealistic Use Cases (Dataset will never be good enough.) 7](#_Toc78399970)

[Methodology 7](#_Toc78399971)

[Initial plan to clean dataset 7](#_Toc78399972)

[Following steps cleaned using OpenRefine. 7](#_Toc78399973)

[Integrity Contraint Violations: DataLog 8](#_Toc78399974)

[Workflows: YesWorkflow 8](#_Toc78399975)

[Future Work 8](#_Toc78399976)

# Introduction & Background

As part of our dedication to open government, transparency and providing high-value data to citizens, USDA had released the Farmers Market Directory listing over a decade ago. The Farmers Market Directory lists markets that feature two or more farm vendors selling agricultural products directly to customers at a common, recurrent physical location. Maintained by the Agricultural Marketing Service, the Directory is designed to provide customers with convenient access to information about farmers market listings to include: market locations, directions, operating times, product offerings, accepted forms of payment, and more. Local farmers markets have proliferated as a means to distribute fresh produce directly to consumers, skipping the costly distribution and packaging step. In this project, we planned to carry out several data cleaning activities which we learnt throughout the course. Few of such activities include: exploring the data, cleaning and standardizing the data, checking integrity violation constraints and producing a final cleaned dataset.

# Data Set

In our data cleaning project, we explore the US Farmers Market dataset from the USDA Website: <https://www.ams.usda.gov/local-food-directories/farmersmarkets>. As defined by Wikipedia, a farmers' market is "a physical retail marketplace intended to sell foods directly by farmers to consumers." The dataset is a directory listing of the various farmers markets in the United States, and includes information such as social media accounts, market location, accepted payments, and agricultural products sold.

# Data Source

Input dataset:

* farmersmarkets.csv - <https://www.ams.usda.gov/local-food-directories/farmersmarkets>

Output dataset(s):

* farmersmarkets\_output.csv
* farmersMarket\_location.csv
* farmeresmarkets\_payments.csv
* farmersmarkets\_products.csv

# Problem Statement

USDA farmers market dataset is a medium sized dataset with some degree of data quality issues. We found a few broad categories of data quality issues. 1. Missing Data 2. Format Issues such as date format 3. Data Type issues such as numeric columns represented as String 4. Data represented in different cases such as upper case, lower case etc., The above mentioned data quality issues pose problems to uniquely identify the entities, locate the addresses and report the various statistics accurately

# Description of Dataset

There are 8687 total observations and 59 columns in this dataset which are described below. The provided html report was generated via a python package called pandas\_profiling, and allows us to observe some basic, preliminary statistics such number of rows and columns, cardinality, missing values, correlations, etc... as well as the overall schema of the dataset.

*FMID* - 7 digit integer that uniquely identifies each farmers' market

*MarketName* - a string containing the name of the farmers' market

*Website, Facebook, Twitter, Youtube, OtherMedia* - a string containing URL or other information that identifies the social media site

*street, city, County, State, zip* - strings that contain data corresponding to the column name that identifies the location of the farmers' market

*Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time* - date fields representing the start date and end date for the given farmers' market or the times in which the farmers' markets are opened

*x, y* - latitude and longitude coordinates

*location* - a string describing the location of the farmers' market

*Credit, WIC, WICcash, SFMNP, SNAP* - Y/N (boolean) character to indicate whether or not a given payment method is accepted

*Organic, Bakedgoods, Cheese...PetFood, Tofu, WildHarvested (30 columns)* - Y/N (boolean) column to indicate whether or not a given product is offered

The following Entity Relationship shows the schema we developed for our dataset. We broke our cleaned dataset into 3 separate tables : location, payments, and products, with the FMID as the primary key for all of them.

Table

Description automatically generated with medium confidence

# Data Exploration

# Data Quality Issues

Using the groups above that describe the dataset contents, we describe some of the quality issues that exist in the dataset, such as non-uniform date formats, Data Type issues such as numeric columns represented as String along with data represented in different cases. We also observe non-uniform null value representation in the same columns.

For the social media columns (Website, Facebook, Twitter, Youtube, OtherMedia), most of the rows appear to be missing, and sometimes, in lieu of an URL, a string is provided.

Table

Description automatically generated with low confidence

The string could be a Facebook username or Twitter handle, but the representation is not uniform. This would directly affect any use case which would involve analyzing social media accounts of farmers market due to missing data. The image below shows some non-null values from the Twitter column, and demonstrates the different types of entries.

Text

Description automatically generated

The location columns that together comprise an address may have some missing values and basically don't contain all 5 components of the address. There may also be leading/trailing white spaces that need to be trimmed, or case conversions that need to be performed, in order to standardize and clean the address data. This directly affects our main use case in the ability to analyze credit card usage by location, which can not be done if the location data is not usable.

Next, for the dates and times, we see that only Season1 tends to be populated. The values are fairly inconsistent as well - some dates are represented using mm/dd/yyyy and some are represented using month name. I've also noticed some date ranges that don't contain the end date. The Season1Time column is also inconsistent. Also, the x and y columns could be better labeled as latitude and longitude, and even the Location column is somewhat poorly because it appears to be a description about the location.

Text

Description automatically generated

Meanwhile, for the boolean columns that contain Y/N values, we also see '-' values which could probably be better represented by a null value. In another words, we want the column to be truly boolean with only 'Y' or 'N'.

A picture containing text

Description automatically generated

Additionally, there are 948 missing values from in the zip code column, about 10.27% of all values. The zip code data is critical to our main use case, so we will scrape data from *tom tom*’s api [ api.tomtom.com], where we can use the latitude and longitude data to obtain the zip code.

Finally, for the updateTime column, we only receive year for some of the records, while others contain the full date time. Also, some of the records contain the month name as opposed to the number.

# Use Case

Given this dataset and our interest in the modernization of payment methods, we think an interesting use case to explore would be identifying the adoption of credit card usage. We could do this by either some SQL queries and in the end, by creating a map that portrays the acceptance of credit cards by state and percentage of markets that accept credit cards.

## Other Potential Use Cases (Dataset "Clean Enough")

*Without (or with very little) additional cleaning, these are just a sample of some of the possible use cases possible with our dataset.*

* We could determine the most and least popular products that tend to be sold by farmers' markets by summing the existence of 'Y' for each product's column. We could also do this across certain states or zip codes. The dataset in its original state would provide enough data to extract this information, we see that the various columns indicating whether or not a certain type of food is sold at a particular market (*columns: Bakedgoods, Cheese, Crafts, Flowers, Eggs, etc.*) are almost entirely populated with very few missing values. Although there is a mix between strings ‘N’ and ‘-‘ for negative values, all positive values are marked ‘Y’; allowing us to sum the total count of ‘Y’ values in each sold item column and rank them to find the most popular options.
* Another use case would be to determine the most popular type of payment options accepted by farmers markets in general (*cash, credit, food stamps, vouchers, etc.*). The dataset in it’s original quality would support this use case because there are very few missing values in the columns indicating payment type accepted; and similar to the sold item columns, all positive cases are marked with a character ‘Y’ which can be used to summarize the metrics of different payment types accepted at farmer’s markets in general.
* We could explore competition within certain zip codes by looking at the density or count of farmers' markets in certain zip codes, due to the original dataset containing very few missing values for the column *zip*.

## Unrealistic Use Cases (Dataset will never be good enough.)

* Detailed analysis of social media options for the farmers markets is also highly unlikely due to missing data. For instance, Youtube, Twitter, and Other Media columns have around 90% missing values. If some of these columns were better populated with links, then a web-scraping pipeline could potentially be developed to augment the current dataset.

# Methodology

## Initial plan to clean dataset

### API client use to fill missing zip code values

We begin by obtaining the missing zip code values for certain columns from tomtom’s reverseGeocode api. [https://api.tomtom.com/search/2/reverseGeocode/{deocode}.json?]

We created a python script to fill in these missing values in the original data [farmersmarkets.csv] and output the updated data set with imputed zip values [farmersmarkets\_imputed.csv].

### Following steps cleaned using OpenRefine.

Next perform some manual data cleaning on the updated data set [farmersmarkets\_imputed.csv] in OpenRefine.

**Step 1.** We begin with the MarketName column by first trimming the leading and trailing whitespace and then collapsing any consecutive whitespaces. We identify key collisions within the MarketName column and cluster and merge these names into common strings.

**Step 2.** Next, we remove some of the columns that are irrelevant to both our main use case and other potential use cases. We have decided that the social media data quality was very poor and so we delete the following columns: Website, Facebook, Twitter, Youtube, OtherMedia. We also remove the time and date columns for Season2 onwards because there is very little data for these.

**Step 3.** Then, we focus on the location columns - street, city, County, State, and zip.

For street, we substitute the ampersand and ‘And’ with 'AND' for uniform representation. Then, we trim the leading and trailing whitespace and collapsed any consecutive whitespaces and then convert to title case.

**Step 4.** We repeat the process of trimming the leading and trailing whitespace and collapsing consecutive whitespaces, along with converting to title case for city, County, and State columns.

**Step 5.** We then check for key collisions on street, city, County and State columns. Both County and State show no key collisions so no further operations are performed on these columns. There are a few key collisions that our found using the fingerprint keying function which appear to be legitimate collisions, so these instances are clustered and corrected. There are also key collisions identified within the street column which are merged.

**Step 4.** Now we move to Season1Date and Season1Time. We decided to just remove these columns because they are not relevant to our current use case. (Of course, we could split Season1Date into 2 columns for a starting and ending date, but we would also have to figure out how we want to represent the rows where a month is given)

**Step 5.** Important to our analysis later, are the x and y columns, which we rename to latitude and longitude respectively, and then convert to numeric. We remove the Location column which is not helpful for our purposes, and is generally blank.

**Step 6.** For some finishing touches, we remove the occurrence of "-" in the Organic column, so that missing values are just left blank.

**Step 7**: We also converted the values in the updateTime column to ISO format using the GREL expression: *value.toDate('d/M/y H:m:s')* after trimming and collapsing whitespace.

### Integrity Contraint Violations: DataLog

### Workflows: YesWorkflow

# Future Work