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

Submission: Team37<Specify NetID: sunilk2, vs24, team\_member3

Abstract:

Table of Content

[Introduction & Background 1](#_Toc76630370)

[Problem Statement 1](#_Toc76630371)

[Data source 2](#_Toc76630372)

[Data Set 2](#_Toc76630373)

[Description of Dataset 3](#_Toc76630374)

[Data Exploration 4](#_Toc76630375)

[Methodology 4](#_Toc76630376)

[Initial plan to clean dataset 4](#_Toc76630377)

[Following steps cleaned using OpenRefine. 4](#_Toc76630378)

[Integrity Contraint Violations: DataLog 5](#_Toc76630379)

[Workflows: YesWorkflow 5](#_Toc76630380)

[Future Work 5](#_Toc76630381)

[Use Case 5](#_Toc76630382)

[Other Potential Use Cases (dataset "clean enough") 5](#_Toc76630383)

[Unrealistic Use Cases (dataset will never be good enough) 6](#_Toc76630384)

# Introduction & Background

# Problem Statement

Using the groups above that describe the dataset contents, we describe some of the quality issues that exist in the dataset from a precursory glance. Many of these will be targeted for cleaning via OpenRefine.

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. The string could be a Facebook username or Twitter handle, but the representation is not uniform. 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.

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.

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'.

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. Again, we will use OpenRefine to correct some of the data quality issues.

# 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

# Data Set

In our data cleaning project, I 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.

# 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

# Methodology

## Initial plan to clean dataset

### Following steps cleaned using OpenRefine.

**Step 1.** We begin with the MarketName column by first trimming the leading and trailing whitespace and then collapsing any consecutive whitespaces. Then we use a text facet and clustering in order to group similar MarketNames together. As seen below, we used the key collision method and the fingerprint keying function.

**Step 2.** Next, we remove some of the columns that are irrelevant to both our main use case and other potential use cases. We had 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 use the following GREL expression to remove any special characters and substitute the ampersand with 'AND', Then, we trim the leading and trailing whitespace and collapsed any consecutive whitespaces and then convert to uppercase.

We go through this exact same process (remove special characters, trim and collapse whitespace, convert to uppercase, clustering) for the city, County, and State columns. After this process, we see that the address information is much cleaner and more consistent.

**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. Because, our analysis is dependent on the Credit column, we make an additional

**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

## Use Case

Given this dataset and my interest in the modernization of payment methods, I 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 (maybe 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.
* We could also determine the most popular type of payment options accepted by farmers markets in general - cash, credit, food stamps or vouchers, etc...
* We could explore competition within certain zip codes by looking at the density or count of farmers' markets in certain zip codes.

## 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 missingness. 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.