

Farmers Market Directory Listing & Payments method Analysis

CS513: Theory *and Practice of Data Cleaning and Provenance* – Final Project (Sum 2021 - Team 37)

Vijayakumar Sitha Mohan(VS24@illinois.edu)

Sunilkumar Kulkarni(SUNILK2@illinois.edu)

Alexander Kokkoros(AMK8@illinois.edu)

Introduction & Overview:

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.

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.

Data Set

In our data cleaning project, we explore the US Farmers Market dataset from the USDA Website and the same can be found at this location:

<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:

We ran data exploration using pandas_profiling and generated basic profiling details such as number of rows, columns, cardinality, missing values, correlations and overall schema details of the dataset.

Total Number of Rows : 8687

Total Number of Columns: 59

To better understand the dataset, we grouped 59 columns into 7 Categories of attributes.

1. FMID - is an identifier for each farmer's market and contains 7 digits.
2. Market name is a free text column.
3. Social Media Sites: Next five columns (Website, Facebook, Twitter, YouTube, OtherMedia) are supposed to contain information that will uniquely identify the market via URI and those are free text columns. All the five columns support blank or null values.
4. Demographics Details : 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.

5. Seasonal Session Details: The following eight columns containing the date and time for every season most likely represent the periods when the market was opened. 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 was held.

6. Market characteristic: The next 35 columns are boolean columns Y/N to represent True/False which represent various indicators such as credit card accepted or not.

7. UpdateTime : The last column (updateTime) representing date and time when the record had been updated

Initial Assessment and Data Quality Issues:

The quick introspection of the dataset suggests there exists the data quality issues, such as Missing Data, Format Issues such as date format, Data Type issues such as numeric columns represented as String and Data represented in different cases such as upper case, lower case etc.,

1. For the social media columns (Website, Facebook, Twitter, Youtube, OtherMedia), most of the rows appear to be missing. It could contain a Facebook username or Twitter handle, but the representation is not uniform.
2. 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.
3. In all the dates and times columns, we could see that only Season1 tends to be populated. We could observe a couple of issues. 1. The date values were not following consistent ISO date formats 2. Also some date ranges that don't contain the end date.
4. The Season1Time column was also inconsistent and did not follow one specific ISO date and time format.
5. Also, the x and y columns could be better labeled as latitude and longitude, and even the Location column appears to be a description about the location.

6. The boolean columns that contain Y/N values and also '-' values which could probably be better represented by a null value.
7. The updateTime column was also not following the consistent format. we could see the column was populated with year only for some of the records and while others contained the full date time. Also, some of the records contained the month name as opposed to the number.

Data Visualization

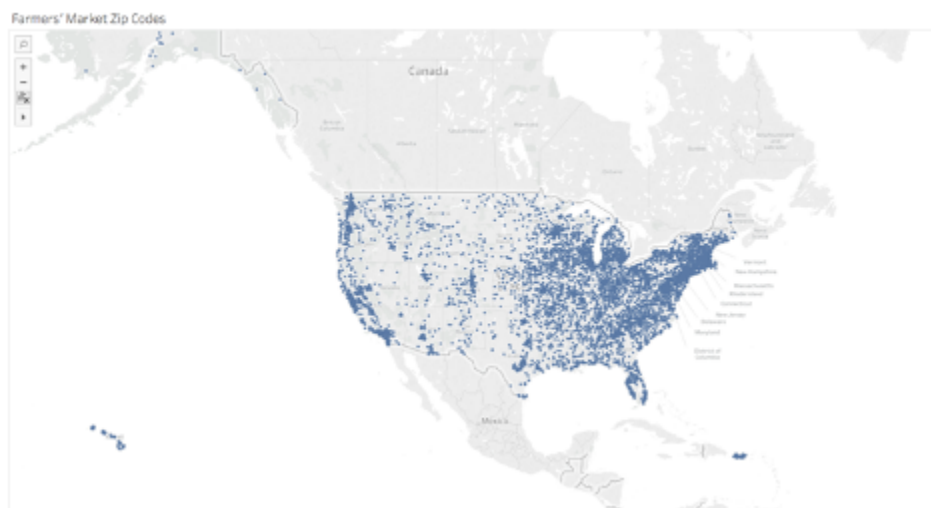
We did quick introspection of the data set and captured a few visuals.

Columns Missing Values Percentage sorted

Column_name	Percentage_missing
y	0.33
x	0.33
city	0.47
street	3.29
County	6.22
zip	10.93
Season1Time	36.35
Season1Date	38.02
Website	39.95
Facebook	56.28
Location	65.90
Twitter	88.52
OtherMedia	92.69
Season2Date	95.04
Season2Time	95.22

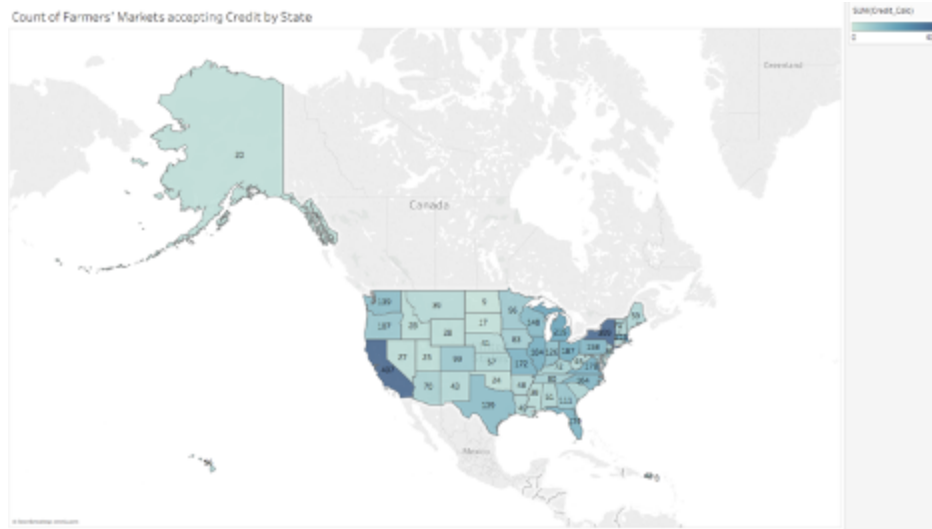
Youtube	98.17
Season3Date	99.09
Season3Time	99.14
Season4Date	99.93
Season4Time	99.93

Farmer Market by Zip Codes



State Wise concentration of Farmer Markets

Next, we see exactly how many farmers' markets each state has, and it is no surprise that more heavily populated states such as California and New York are shaded darker.



Use Cases:

Is data clean enough for use cases?

Use Case : Farmers Market Directory Listing

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

- Users can query the listing by name, location, by product etc., interactively or through API.
- 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.
- In what period of time the farmers market was opened the most or least. To do so, season1 date should be reformatted in the same format such as DD/MM/YYYY.

Hypothetical use case

Use Case : E-Payment Adoption

After cleaning and standardizing the dataset, we could explore different payment methods supported by vendors, comparison among the vendors by supported payment methods and which are popular payment methods versus least favorite.

With additional dataset, We could combine our farmers market demographic data with US counties income distribution data to identify the right location to host the farmer markets so that the business can grow for farmers and for citizens, they could get organic and healthy produce.

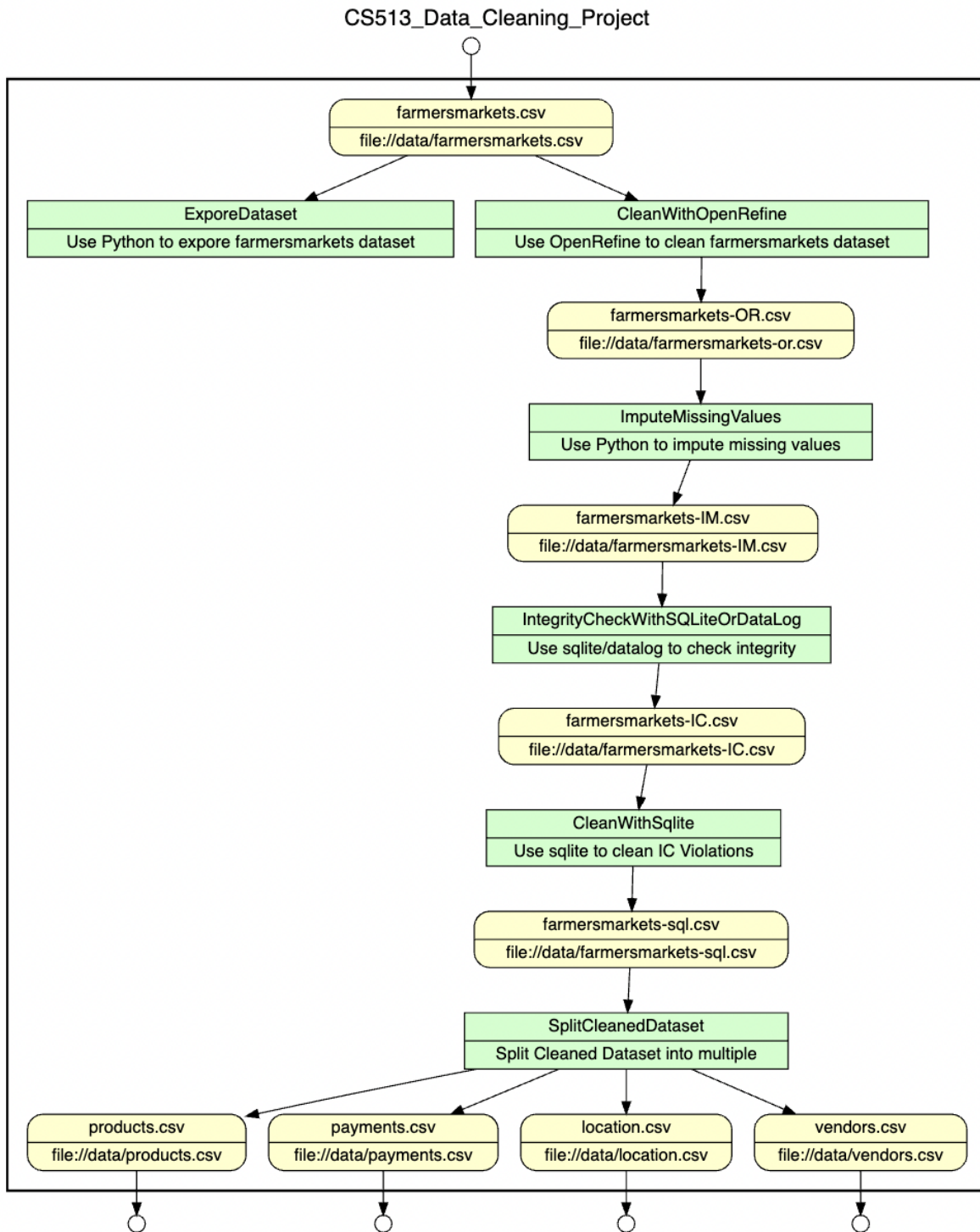
With additional dataset such as sales data, we could analyze the payments and suggest the vendors to adopt the modern payment methods instead of cash.

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.
- Season3 and Season4 date and time were missing for above 90% of the records so we were not able to perform any analysis based on date time for those seasons.

Methodology

We have captured the intended Data Cleaning and Provenance Steps in the YesWorkflow.



OpenRefine Data Cleaning Steps

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 deleted 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 collapse 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 Constraint Violations: DataLog/Sqlite

We will be performing Integrity constraints check using DataLog or Sqlite

1. To uniquely identify Farmer Market Vendors
2. To clean and map the vendors to exact location

Workflows: YesWorkflow

We will be capturing the detailed project steps and provenance using YesWorkflow. The initial sample YesWorkflow diagram included above.

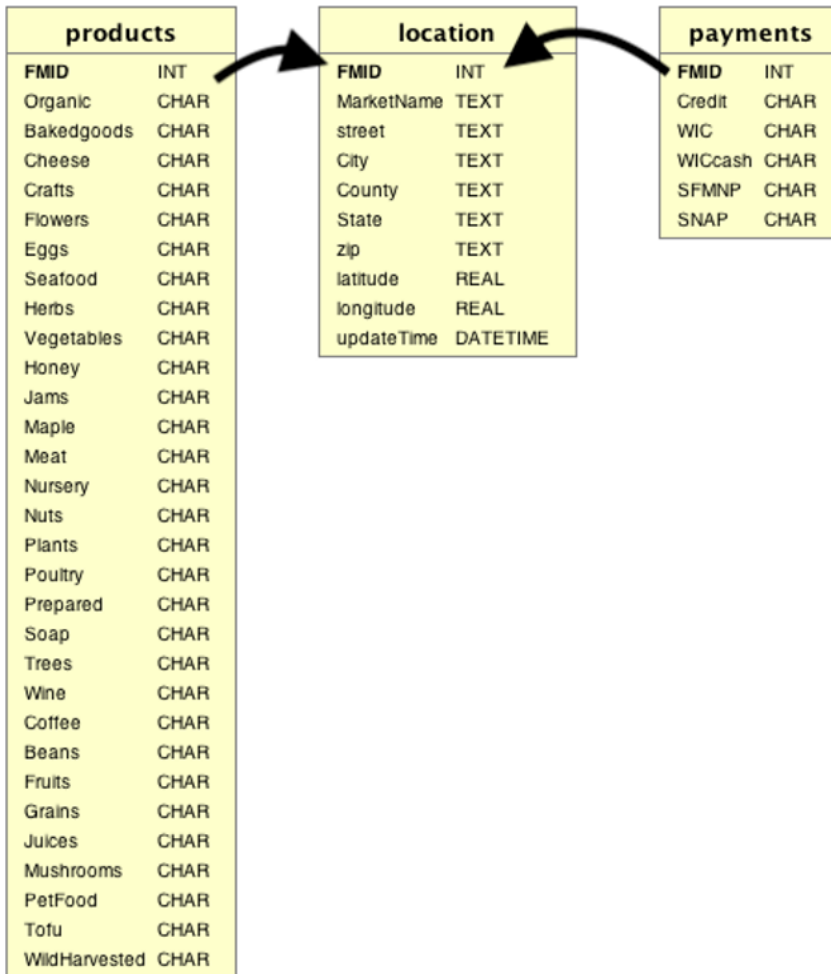
Results:

After applying the data cleaning, wrangling activities, we planned to capture the below datasets and their relationships as shown below.

Output dataset(s):

- farmersmarkets_vendor.csv
- farmersMarket_location.csv
- farmeresmarkets_payments.csv
- farmersmarkets_products.csv

The following Entity Relationship shows the schema we developed for our dataset. We broke our cleaned dataset into 4 separate tables : `vendor`, `location`, `payments`, and `products`, with the FMID as the primary key for all of them.



Conclusion

After performing manual cleaning using OpenRefine, Used DataLog/Sqlite to check integrity constraints and curation and then able to create desired output tables. Then we use these tables to analyze the use case in interest. We were able to generate a few tableau visualizations to see farmer market concentration in the US by zip code and state wise.

The first one plots the zip codes corresponding to the farmers' markets, and it is essentially like a density plot that allows us to see that some of the more densely populated regions, for example in the northeast, have many farmers' markets, while the midwest and Alaska appears to be sparser.

Also create few visualizations for payment adoptions and popularity.

References

US Department of Agriculture

(<https://www.ams.usda.gov/local-food-directories/farmersmarkets>)

Kaggle

(<https://www.kaggle.com/madeleineferguson/farmers-markets-in-the-united-states/activity>)