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 Exploration

Initial Data Analysis and Exploration:

We performed initial data analysis using python in Jupyter notebook. We looked into the following aspects with the dataset.

1. Fill rate for each attribute
2. Data types of each attribute
3. Missing values of certain interested attributes

The following table shows the percentage of missing values of each column in the dataset, used to identify unusable columns and information.

```
Table of Missing percentages for each attribute
FMID - 0%
MarketName - 0%
Website - 40%
Facebook - 56%
Twitter - 89%
Youtube - 98%
OtherMedia - 93%
street - 3%
city - 0%
County - 6%
State - 0%
zip - 11%
Season1Date - 38%
Season1Time - 36%
Season2Date - 95%
Season2Time - 95%
Season3Date - 99%
Season3Time - 99%
Season4Date - 100%
Season4Time - 100%
x - 0%
y - 0%
```

```
Location - 66%
Credit - 0%
WIC - 0%
WICcash - 0%
SFMNP - 0%
SNAP - 0%
Organic - 0%
Bakedgoods - 0%
Cheese - 0%
Crafts - 0%
Flowers - 0%
Eggs - 0%
Seafood - 0%
Herbs - 0%
Vegetables - 0%
Honey - 0%
Jams - 0%
Maple - 0%
Meat - 0%
Nursery - 0%
Nuts - 0%
Plants - 0%
Poultry - 0%
Prepared - 0%
Soap - 0%
Trees - 0%
Wine - 0%
Coffee - 0%
Beans - 0%
Fruits - 0%
Grains - 0%
Juices - 0%
Mushrooms - 0%
PetFood - 0%
Tofu - 0%
WildHarvested - 0%
updateTime - 0%
```

We see in this table that most of the food and items for sale are completely filled, as well as the update time, payment methods, coordinates, FMID, market name, city and state. We see that the social media sites are mostly missing and are likely unusable for general data analysis, as well as the season 2-4 dates and times. We see that an attribute of interest, zip has 11% of values missing which will need to be addressed in order for future analysis. We wanted the location such as Zip code and City to be available 100%. As you see from the above table zip was missing for ~ 11% and city was missing for about 0.47% records. So, we decided to impute the values for zip and city based on other available attributes.

We also examine the number of unique values for our attributes in the table below to see if they fit with our understanding of the data.

Number of Unique Values

```
FMID             8675
MarketName       8102
Website          4273
Facebook         3347
Twitter          748
Youtube          122
OtherMedia       495
street           8196
city             5012
County           1490
State            53
zip              6277
Season1Date      2359
Season1Time      1700
Season2Date      378
Season2Time       205
Season3Date      77
Season3Time      45
Season4Date      6
Season4Time      6
x                8533
y                8533
Location         10
Credit           2
WIC              2
WICcash          2
SFMNP            2
SNAP             2
Organic          3
Bakedgoods       2
Cheese           2
Crafts           2
Flowers          2
Eggs             2
Seafood          2
Herbs            2
Vegetables       2
Honey            2
Jams             2
Maple            2
Meat             2
Nursery          2
Nuts             2
Plants           2
Poultry          2
Prepared         2
Soap             2
Trees            2
Wine             2
Coffee           2
Beans            2
Fruits           2
Grains           2
Juices           2
```

```
Mushrooms          2
PetFood            2
Tofu               2
WildHarvested      2
updateTime         6154
```

We notice here, that FMID has 8675 unique values, and there are 8675 rows in this data set which leads us to believe that FMID will serve as a primary key in its current state, this will be confirmed later with SQLite integrity constraint checking. We also notice that MarketName has 8102 unique values indicating that some markets share a name, and this attribute cannot be used as a key. We also notice here that some attributes which should be binary, such as Organic have 3 values, instead of 2, which may indicate non-uniform representation of null values and should be explored further.

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.
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.

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 cannot 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.

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'.

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 a 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.

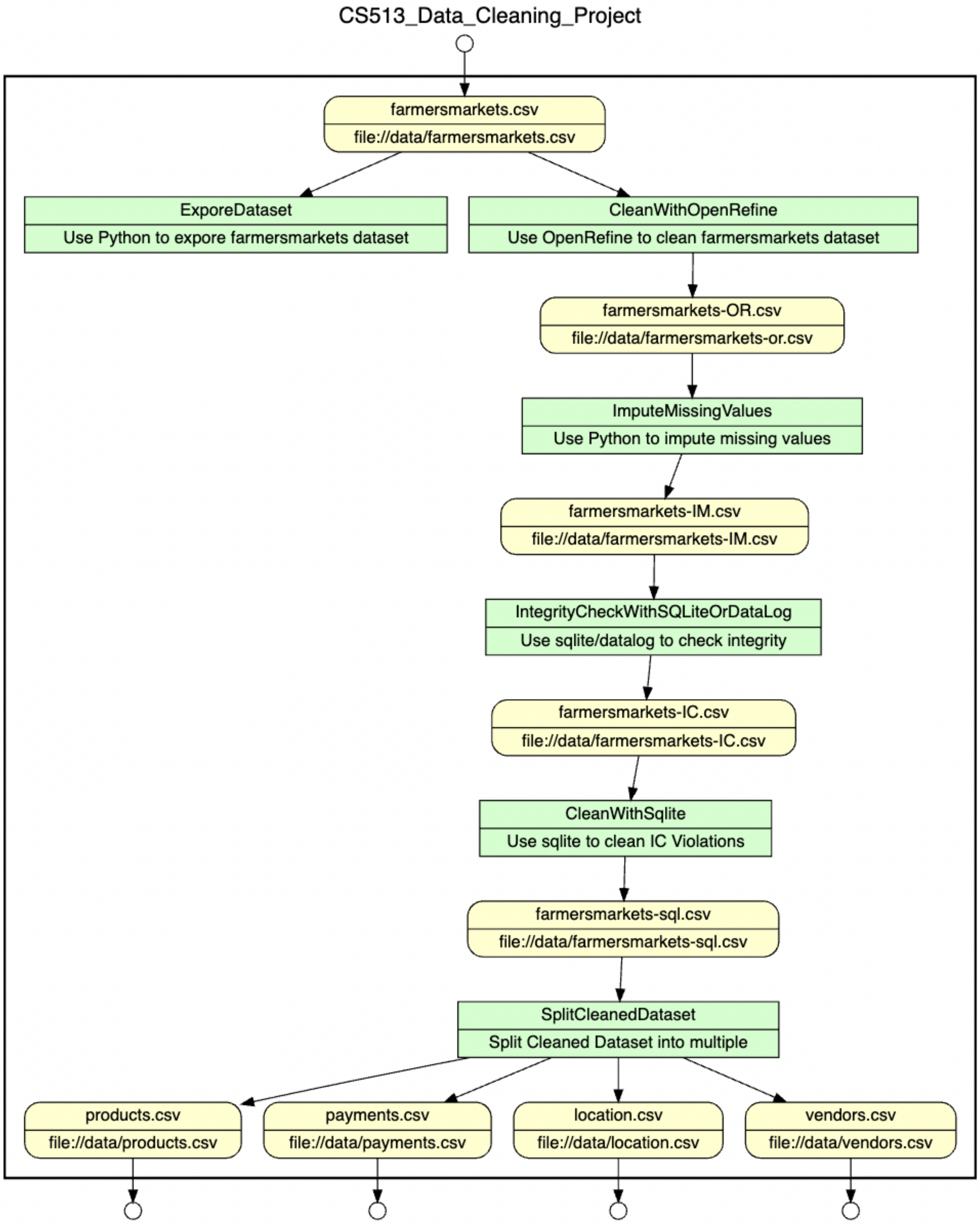
- We could explore competition within certain zip codes by looking at the density or count of farmers' markets in certain zip codes, due to the original dataset containing very few missing values for the column *zip*.

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

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


CS513_Data_Cleaning_Project
farmersmarkets.csv
file://data/farmersmarkets.csv
ExporeDataset
Use Python to expore farmersmarkets dataset
CleanWithOpenRefine
Use OpenRefine to clean farmersmarkets dataset
farmersmarkets-OR.csv
file://data/farmersmarkets-or.csv
ImputeMissingValues
Use Python to impute missing values
farmersmarkets-IM.csv
file://data/farmersmarkets-IM.csv
IntegrityCheckWithSQLiteOrDataLog
Use sqlite/datalog to check integrity
farmersmarkets-IC.csv
file://data/farmersmarkets-IC.csv
CleanWithSqlite
Use sqlite to clean IC Violations
farmersmarkets-sql.csv
file://data/farmersmarkets-sql.csv
SplitCleanedDataset
Split Cleaned Dataset into multiple
products.csv
file://data/products.csv
payments.csv
file://data/payments.csv
location.csv
file://data/location.csv
vendors.csv
file://data/vendors.csv

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.

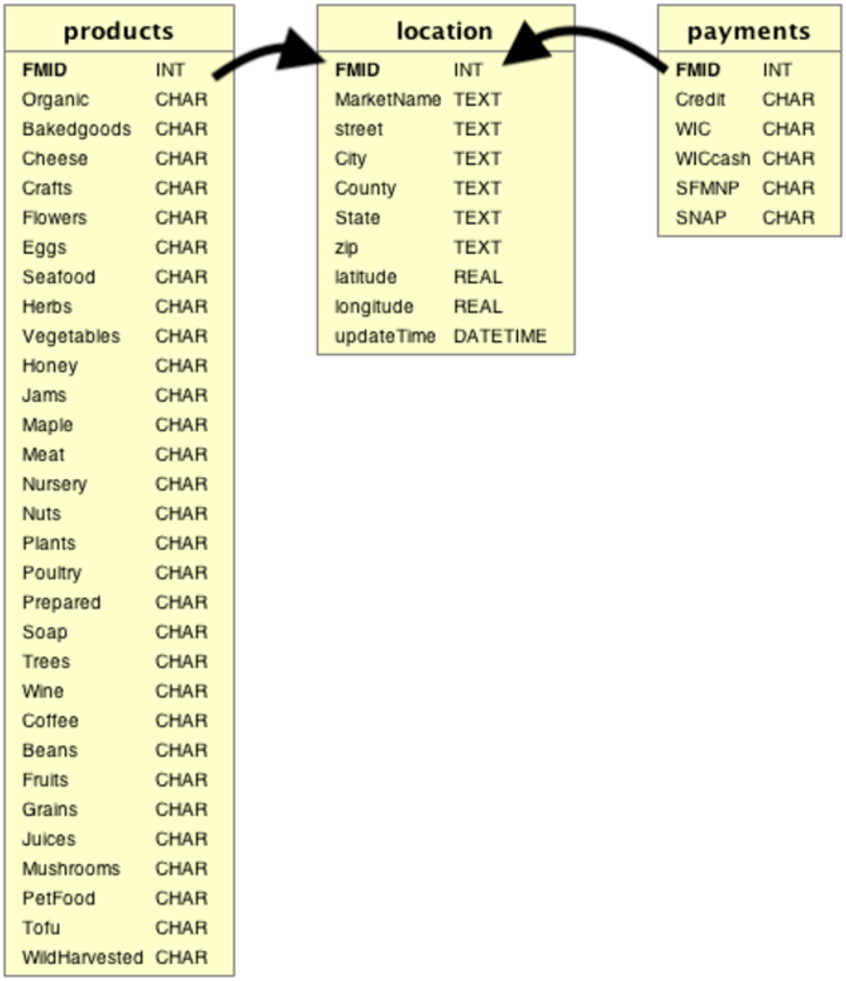
Results:

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.



products	
FMID	INT
Organic	CHAR
Bakedgoods	CHAR
Cheese	CHAR
Crafts	CHAR
Flowers	CHAR
Eggs	CHAR
Seafood	CHAR
Herbs	CHAR
Vegetables	CHAR
Honey	CHAR
Jams	CHAR
Maple	CHAR
Meat	CHAR
Nursery	CHAR
Nuts	CHAR
Plants	CHAR
Poultry	CHAR
Prepared	CHAR
Soap	CHAR
Trees	CHAR
Wine	CHAR
Coffee	CHAR
Beans	CHAR
Fruits	CHAR
Grains	CHAR
Juices	CHAR
Mushrooms	CHAR
PetFood	CHAR
Tofu	CHAR
WildHarvested	CHAR

location	
FMID	INT
MarketName	TEXT
street	TEXT
City	TEXT
County	TEXT
State	TEXT
zip	TEXT
latitude	REAL
longitude	REAL
updateTime	DATETIME

payments	
FMID	INT
Credit	CHAR
WIC	CHAR
WICcash	CHAR
SFMNP	CHAR
SNAP	CHAR

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)