Project Phase-I

Identify a dataset

For this project, we will utilize the **Farmer's Market Dataset.** This dataset contains a number of farmer's markets along with data such as their name, location, open times, payment methods accepted, and products sold.

Use Cases

Before we embark on cleaning the data, we must identify a purpose for which the data needs to be clean. In this section, we will identify three use cases: target use case (U1), "zero data cleaning" (U0) use case, and "never (good) enough" (U2) use case.

- Target use case (U1):
  - "Which Farmer's Markets sell vegetables, meats, *and* seafood?"
    - This question would help a customer find the appropriate farmer's markets that sell vegetables, meats, and seafood to help plan their grocery trip.
  - "Which Farmer's Markets sell organic produce?"
    - This use case would help the customer find appropriate farmer's markets that sell organic produce
  - "Could I map the density of Farmer's Market which accept Credit/WIC/WICcash/SFMNAP/SNAP across the U.S.?
    - With almost complete data found in location and products portion of the dataset, it would be interesting to see the question mapped out across the Farmer's Market in the U.S.

- "Zero Data Cleaning" Use Case (U0):
  - "Which Farmer's Market accepts Credit, Cash, WIC, WICash, SFMNP, and/or SNAP?"
    - As the above attribute values are all filled, no data cleaning is required to answer this use case.
  - "Can I create a map of farmer's markets by city, state, and county to get an understanding of the density of the farmer's market in the U.S?"
    - As *Street*, *City*, *County*, and *State* information is available to a large extent, a user would be able to get a good idea of this question without any data cleaning. Also, by using the *x* and *y* coordinate fields, this question would be easily and accurately answered.
  - "Can I determine the popularity of X product in Y State/City/Zip?"

        - As the product attribute values are abundantly available in the Farmer's Market Dataset, this question could be answered using the given information.

- "Never Good Enough" Use Case (U2):
    - "Which farmer's market can I visit on X date?"
        - The Farmer's Market Dataset is very limited and poorly displays the dates and times that the markets are open and available for a visit. *Season1Date* attribute is mostly filled with date ranges and Season1Time is filled with weekdays and times the said market is open. Though, *Season2Date*, *Season2Time*, *Season3Date*, *Season3Time*, *Season4Date*, and *Season4Time* are either missing data or display inaccurate or redundant data. Moreover, the seasons do not accurately represent the four seasons. For example, in the United States, Summer lasts from June 21st to September 22nd. If we presume *Season1Date* to be the summer season, most date ranges do not adhere to this range. With 8666 total Farmer's Market, it would be impossible to sort out the data.
    - "Could I scrape the social media URLs for information about the Farmer's Market?"
        - As most of the URLs for the social media sites are missing, non-URL strings, or no longer exist, this question would not yield useful results.

Describe the Dataset

The Farmer's Market Dataset contains 59 columns and 8666 rows. The attribute name *FMID* is a primary key (PK) that uniquely identifies each record. Followed by the attribute name *MarketName*, which identifies each farmer's market. The next five field names are dedicated to social media URLs, which are filled with inconsistent and missing data. Followed by locational data such as *street*, *city*, *County*, *State*, and *Zip*. These attributes are all nearly filled, though require various forms of data cleaning. Next, we have eight seasonal data attributes composed of *Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, and Season4Time*; apart from Season1 date and times most of the other attributes have little to no information. Followed by *x* (Longitude) and *y* (Latitude) coordinates, which have near-complete data, and ideally, the attributes would be moved next to location information. Next, we have the *Location* field, which contains miscellaneous location-based information about it few of the farmer's markets. Next we information about payment methods accepted by the farmer's markets. *Credit, WIC* (Women, Infants, Children Program), *WICcash* (Women, Infants, Children Cash Program), *SFMNP* (Seniors Farmers' Market Nutrition Program), and *SNAP* (Supplemental Nutrition Assistance Program). The data is noted as Y/N. These attributes are also completely filled with no missing values. Followed by the products sold at the farmer's markets: *Organic, Bakedgoods, Cheese, Crafts, Flowers, Eggs, Seafood, Herbs, Vegetables, Honey, Jams, Maple, Meat, Nursery, Nuts, Plants, Poultry, Prepared, Soap, Trees, Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu,* and *WildHarvested.* These attributes are missing a large number of data values, but a good understanding can be gathered from the values present. At last, we have the *updateTime* attribute, which can presumably be described when the data record was last updated for each of the Farmer's Markets.

As social media data does not provide any important relevancy to the use case U1, we have removed the attributes from the dataset.

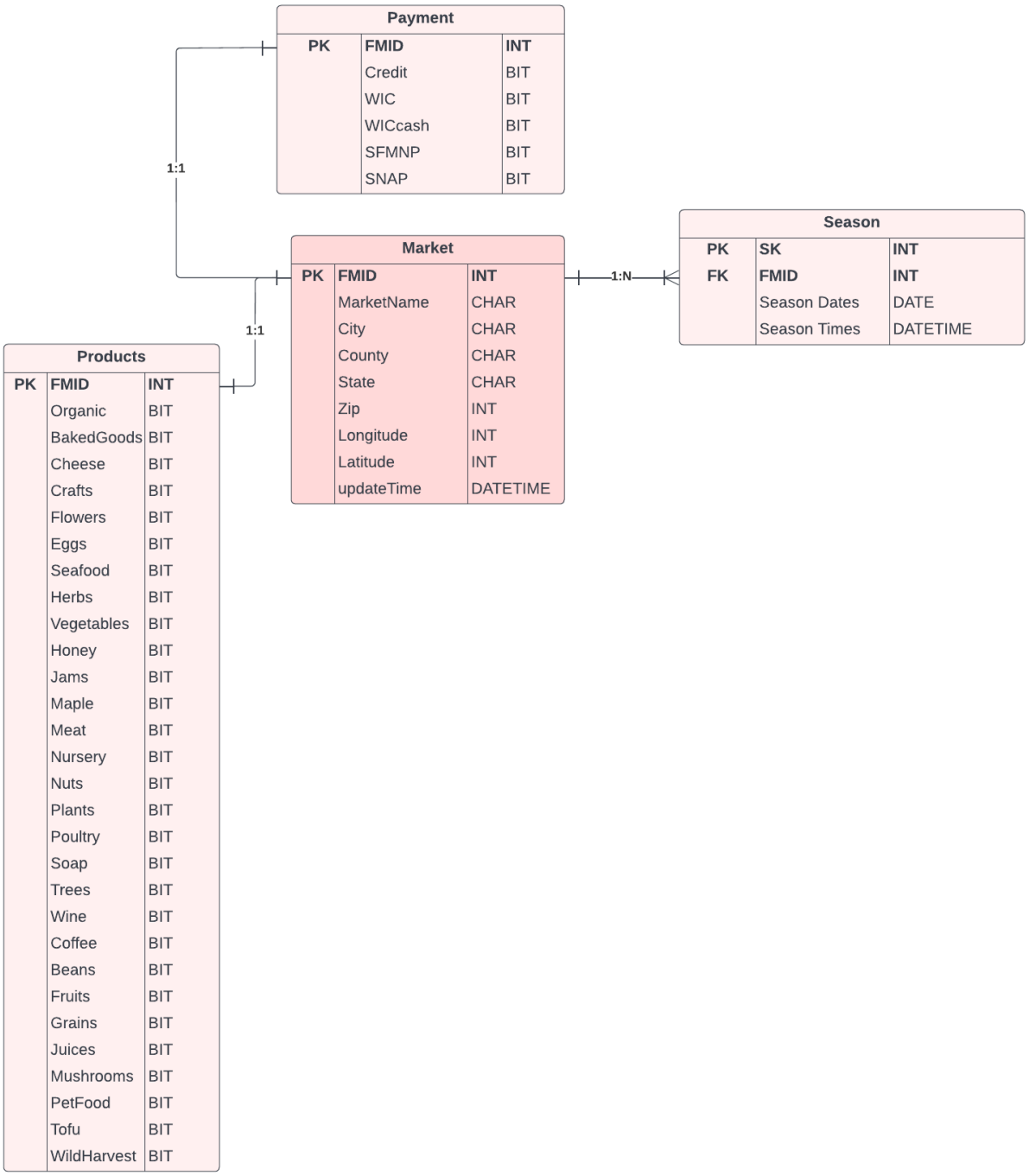

Figure 1 Database Schema for Farmer's Market Database

Data Quality Problems

At an initial glance, we can notice several data quality issues present in the Farmer's Market Dataset.

The goal is to achieve data quality sufficient to fulfill U1. To achieve this, we must focus on reducing errors and improving the consistency of the farmer's market names, location data, accepted payment information, and product data. Though in general proper data cleaning exercises will improve the accuracy, consistency, and reliability of all attributes.

Using the text facet in OpenRefine, we can observe that farmer's markets, street addresses have inconsistencies in title cases and spellings. The same is consistent with other attributes such as city and county. We must canonicalize the data here using clustering in OpenRefine.

Figure 2 MarketName cluster in OpenRefine

Figure 3 Street Cluster in OpenRefine

Social media attributes (*Website, Facebook, Twitter, Youtube, OtherMedia*) do not provide any important relevancy to the use case U1, hence it is logical to remove the attributes from the dataset.

It is also vital to rename some of the attributes to improve consistency. For instance, *x* and *y* attributes would become *Longitude* and *Latitude*, respectively. This includes attribute names that need a proper title case and spell check for persistency with the rest of the dataset.

x	y
-72.140305	44.411013
-81.728597	41.375118
-85.574887	42.296024
-82.8187	34.8042
-94.274619	37.495628
-73.9493	40.7939

Longitude	Latitude
-72.140305	44.411013
-81.728597	41.375118
-85.574887	42.296024
-82.8187	34.8042
-94.274619	37.495628
-73.9493	40.7939

Figure 4 x and y transformation to Longitude and Latitude

It is important to observe the social media attributes (*Website, Facebook, Twitter, Youtube, and OtherMedia*) as they have a vast number of missing and inaccurate information and it is not contributing to our goal U1 use case. It must also be noted that the Location attribute does not provide any significant use to our U1 goal, hence it can be removed from the dataset as well.

As SeasonXDate and SeasonXTime do not follow the date ranges of the four seasons, it is not prudent to allocate them eight individual attributes. It will be simpler for comprehension to pull this data into a separate table and combine the date attributes into a single attribute, Season Date, which contains the date ranges that the farmer's market is open; along with a single attribute time, which combines the times. Please see the detailed steps in Initial Plan.

It can also be observed that the attributes are not correct data types. It will be vital to add this to the plan to correct it. The attributes that contain date and time information will need to be made ISO compliant standard date form YYYY-MM-DD

A case can be made to convert Y and N, a CHAR data type into a BIT data type, found in payment attributes and product attributes. Even though CHAR(1) and BIT data type only consume 1 byte to store, when large data is considered in SQL server, we can add up to 8 columns of a BIT into a single byte, while each column of type CHAR(1) will take up one byte. This will optimize space in the case our dataset continues to grow.

Some zip codes appear to be only 4 characters. We must apply integrity constraints to confirm that all zip codes have 5 characters.

Initial Plan

S1: Description of dataset D and matching use case U1

The Farmer's Market Dataset contains 59 columns and 8666 rows. The matching use case(s) we will explore is defined as the following:

- "Which Farmer's Markets sell vegetables, meats, and seafood?"
    - This question would help a customer find the appropriate farmer's markets that sell vegetables, meats, and seafood to help plan their grocery trip.
- "Which Farmer's Markets sell organic produce?"
    - This use case would help the customer find appropriate farmer's markets that sell organic produce
- "Could I map the density of Farmer's Market which accept Credit/WIC/WICcash/SFMNAP/SNAP across the U.S.?"
    - With almost complete data found in location and products portion of the dataset, it would be interesting to see the question mapped out across the Farmer's Market in the U.S.

The attributes of D are specified as below:

Attribute Name	**Description**
FMID	Primary Key that uniquely identifies each record
MarketName	Name of the Farmer's Market
Social Media (Website, Facebook, Twitter, Youtube, OtherMedia)	URLs to the social media of the Farmer's Market
street	Street address of the Farmer's Market
city	City of the Farmer's Market
County	County of the Farmer's Market
State	State of the Farmer's Market
zip	Zip code of the Farmer's Market
Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time,	Date ranges and time frames that the Farmer's Market is open
x	Longitude
y	Latitude
Location	Miscellaneous location information for Farmer's Market

Credit, WIC, WICcash, SFMNP, SNAP	Payment methods accepted
Organic, Bakedgoods, Cheese, Crafts, Flowers, Eggs, Seafood, Herbs, Vegetables, Honey, Jams, Maple, Meat, Nursery, Nuts, Plants, Poultry, Prepared, Soap, Trees, Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, and WildHarvested	Products sold at the Farmer's Market
updateTime	The last known date/time the record was updated

S2: Profiling of D to identify the quality problems P that need to be addressed to support U1

Along with the data quality problems are defined in Data Quality Problems, it is critical when profiling a dataset to identify natural groups and characteristics that can help classify the dataset into meaningful and feasible relations. Based on our analysis, this dataset breaks down into the following natural groups:

- Market Data: *FMID (PK), MarketName, City County, State, Zip, Longitude, Latitude, updateTime*
- Products: *FMID (PK), Organic, Bakedgoods, Cheese, Crafts, Flowers, Eggs, Seafood, Herbs, Vegetables, Honey, Jams, Maple, Meat, Nursery, Nuts, Plants, Poultry, Prepared, Soap, Trees, Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, and WildHarvested*
- Payment: *FMID (PK), Credit, WIC, WICcash, SFMNP, SNAP*
- Season: *SK (PK), FMID (FK), Season Dates, Season Times*
    - We have decided to combine the season dates and times into one date and one time attribute, we have created a separate table to keep track of them. As one Farmer's Market can have multiple dates/times that they are open, the Market Data and Season tables have a one-to-many relationship.
    - Note: SK stands for Season Key, and FK stands for Foreign Key

Please refer to the subsection Data Quality Problems and Figure 1 for details here.

S3: Performing the data cleaning process using one or more tools to address the problems P

Here are the tools we will utilize to clean the Farmer's Market Dataset. Please note that this list is subject to change as more tools are introduced in the class:

- OpenRefine:
    - Get an overview of the dataset
    - Remove minor inconsistencies such as removing whitespaces and trailing whitespaces
    - Resolve inconsistencies in the dataset
    - Apply integrity constraints

    - Using Facets and Clustering to canonicalize the dataset
    - Filter data using specific parameters
- Python:
    - Use Dataframe to merge Season Date and Time Columns
- SQL:
    - Create a new database with the relations shown in the database schema
    - Run queries to confirm U1 goals can be met
- Jupyter Notebook:
    - Visualizing and analyzing the data
    - Map the U1 use case regarding payment usage

S4: Checking that your new dataset D' is an improved version of D

By utilizing the data cleaning methods learned in this class, the new dataset D' will be rid of errors found in D, such as inconsistent and inaccurate data. D' will be clear, concise, and easier to query the U1 use case goal. As the previous dataset D was not able to be queried due to data types inconsistencies and non-canonicalized data, dataset D' will not present these challenges due to the efforts we have made.

S5: Documenting the types and number of changes that have been executed on D to obtain D'

Using OpenRefine, we will need to perform the following steps to initially clean the data:

1. Remove attributes not relevant to the U1 goal
    - This includes social media attributes: *Website, Facebook, Twitter, Youtube, OtherMedia*
    - *Location*
2. Remove any unnecessary whitespaces, including trimming them from the beginning and end of all data
3. Correct data types:
    - Text: *MarketName, street, city, County, State, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time*
    - Number: *FMID, zip, x, y*
4. Rename to attributes for consistency:
    - *x* to *Longitude*
    - *y* to *Latitude*
    - *street* to *Street*
    - *city* to *City*
    - *zip* to *Zip code*
    - *Bakedgoods* to *BakedGoods*
5. Make the following attributes title case:
    - *Street*
    - *City*
    - *County*

    - *State*
6. Convert the following attributes from Y to 1 and N to 0:
    - Credit, WIC, WICcash, SFMNP, SNAP
    - Organic, Bakedgoods, Cheese, Crafts, Flowers, Eggs, Seafood, Herbs, Vegetables, Honey, Jams, Maple, Meat, Nursery, Nuts, Plants, Poultry, Prepared, Soap, Trees, Wine, Coffee, Beans, Fruits, Grains, Juices, Mushrooms, PetFood, Tofu, and WildHarvested
7. Use text facets and clustering to clean the following attributes:
    - *MarketName*
    - *Street*
    - *City*
    - *County*
    - *State*
8. Transform the date attributes to the ISO-compliant YYYY-MM-DD:
    - *updateTime*
9. Apply integrity constraints and remove zip codes with 4 characters

Python:

1. Confirm all *FMIDs* are unique
2. Combine SeasonXDate attributes and SeasonXTime into Season Date and Season Time respectively.

SQL:

1. Export the data and use SQL to separate the dataset into its natural groupings.
2. Confirm the data types conform to the ones specified in the database schema

Jupyter:

1. Export the data and use python libraries to visualize the data