CS513 Data Cleaning Project - Phase 1

Ehab Aleid ealeid2@illinois.edu

Diana Arita dianama2@illinois.edu

Kevin Zhong Kzhong7@illinois.edu

7/10/2021

**1. Identify a dataset**

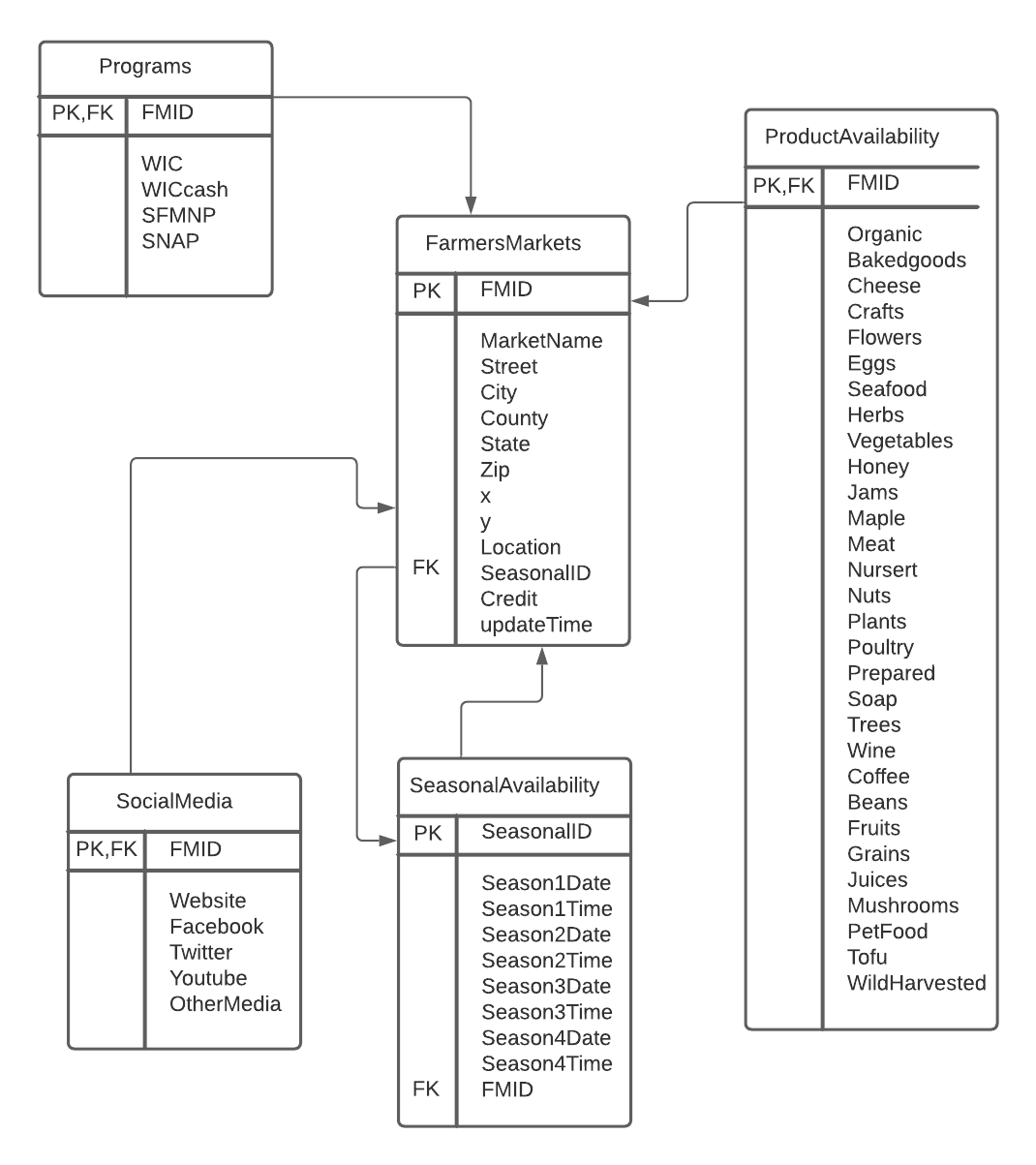
For this project we chose the US Farmers Markets for our dataset D.

**2. Target Use Case**

* The target use case (U1) will be to create a website that will allow users to search for a Farmers website based on the users specific search criteria such as location, products, and payment method accepted . After performing a search, the user will see the official website of the farmers market along with any other social media links. In today’s modern world, every business needs to have a website. Websites help organizations , such as farmers markets reach out to new consumers. However, within the dataset, some of the websites are incorrectly formatted, missing, or have shortened URLs.
* A simple use case (U0) is to simply allow users to search a farmers market by name or city, and see what is sold at the market. However, this use case is too simple, since the dataset already provides the name and crops sold at the market.
* A more complicated use case (U2) is to also provide the address of the farmers market. This can't be fixed because some of the data has the address listed as "HW 2" or "Route 27". Also some of these farmers markets may have a different location every year. Since U1 will provide the website and other social media platforms of each farmers market, a user should be able to find the location of the farmers market from the website

**3. Dataset Description**

* We created a relational schema for the farmers markets dataset. The schema contains 5 tables. The first table is titled ‘FarmersMarkets’ and has 13 columns. These columns describe the farmers markets name, location, credit acceptance, and updated time. The table has a foreign key pointing to its seasonal ID. The second table has the title ‘Programs’, it’s six columns provide details on whether a farmers market accepts certain programs. The third table is titled ‘ProductAvailability’ and contains columns for different products that are sold and whether or not the farmers markets sell it. The fourth table is titled ‘SocialMedia’ and has columns for certain social media platforms. If a farmers market uses the social media market then it’s link will be in its cell. The fifth and last table is titled ‘SeasonalAvailability’. It’s columns show which seasons the farmers markets are available. The schema contains data for the United States. Based on the updatedTime column in the ‘FarmersMarkets’ table, the period for the data is 2010 to 2017.



**4. Data Quality Problems**

When reviewing the farmers market dataset we quickly find many data quality problems that need to be cleaned to enable our use case U1. For example, many of the markets have repeated records in multiple rows. This dataset needs to be normalized to reduce redundancy and improve data integrity. Another data quality issue we find is incorrect/inconsistent data values for some records. Some examples of these data quality issues can be seen for the farmers market “Alexandria Farmers Market” which has multiple records for the same market with unique FMID’s. In addition to having multiple records for a single farmers market, one of the records incorrectly lists a facebook link in the website URL attribute and also incorrectly has the market name in the facebook link attribute. Furthermore we find many instances of an attribute having record values with numerous types of format. The values of each attribute should follow a similar encoding/format to allow ease of data usability for our use case. An example of this is in the “Season1Date '' attribute, some records are formatted in a numeric MM/DD/Year format and others have a string “Month\_X to Month\_Y” format. This needs to be solved so our use case has consistent formatting for the website. Finally we also find many instances of missing data and inconsistent syntax for the same strings which need to be resolved.

**5. Initial Plan**

S1 - Description of the dataset and its matching use case

* The dataset can be described using a relational schema. We will simplify the data to 5 tables: FarmersMarkets, Programs, ProductAvailability, SocialMedia, and SeasonalAvailability. The data will contain different types including varchar, bit, and datetime. The farmers markets are all in the United States and the data time range is between 2010 and 2017.
* The target use case (U1) will be to create a website that will allow users to search for a Farmers website based on the users specific search criteria such as location, products, and payment method accepted . After performing a search, the user will see the official website of the farmers market along with any other social media links.

S2 - Data Quality Problems

* The dataset has many quality problems that need to be resolved with Open Refine and other tools to achieve U1. Firstly the dataset needs to be normalized such that redundant records are merged into a single record for each unique farmers market. Secondly invalid data cell values need to be scrubbed or updated with a consistent format. Thirdly, syntax errors need to be cleaned so that similar strings use the same spelling.

S3 - Steps to perform the data cleaning process

* Open Refine will be the main tool used to clean dataset D.
* Use Open Refine to merge redundant records
* Invalid data cells will need to be removed or updated to valid format
* Dates will be reformatted to a consistent format.
* Syntax errors will be corrected

S4 - Checking the new dataset is an improved version.

* SQL will be used to confirm that the cleaned dataset does not contain the data quality problems we defined.

S5 - Documenting the types and amount of changes that have been executed.

* Open Refine History
* The outer workflow will be created using YesWorkflow
* The inner workflow for the Open Refine history will be created using Yes Workflow and the or2yw tool.

**Tentative Assignments of Tasks**

* Ehab Aleid will use Open Refine to clean the dataset
* Diana Arita will use SQL to check that the new dataset D’ is an improved version of D and D follows integrity constraint rules.
* Kevin Zhong will model the workflow using YesWorkflow.