CS513: Farmers Markets

**Final Project Report (Summer 2022) - Phase 1**

| Dinesh Agarwal | Sahib Singh | Rahul Sharma |
| --- | --- | --- |
| dinesha3@illinois.edu | fixme@illinois.edu | rahul9@illinois.edu |
| UIUC | UIUC | UIUC |

1. **Dataset (D)**

USDA experts gather this data from assistance, conservation, and recovery programs for agricultural producers. As a data-driven organization, USDA uses this data to manage and implement programs – and to plan for the future.

We found this dataset to be most suitable for our project by analyzing it based on following criterias fidelity, high cardinality and size. Moreover, we are using various interactive data visualization and exploration tools for our project.

1. **Use Cases (U)**

* **U0(which do not require data cleaning):** Identity the states which are most popular for farming?

We came up with this use case where we would plot the density of numbers of Farmer’s Market per state across the U.S as a heat map. In this heat map, each state color becomes more green as the number of farmer’s markets in that state increases.   
We leverage farmers’ demographics i.e. latitude, longitude & city name which does not require any data cleaning since the data is “good enough as it is”.

**SELECT city,**

***count*(*\**) as cnt**

**FROM farmers\_market**

**WHERE LAT <> NULL**

**OR LON <> NULL**

**GROUP BY city**

**ORDER BY cnt DESC;**

* **U1(the one require some cleaning and can be achieved):**

1. What is the peak season during which most of the Farmer's Market are open?We observed that Seasonal Date and Time columns have unclean data. Primarily they have different formats in which date and timings are reported. We believe that we can normalize the various time dimensions in the dataset and hence clean it to solve this use case.

1. Remove redundant information from the dataset.

We see data redundancy issues across demographics about Farmer’s Market and with the help of RDBMS normalization concept we can make it easy to update address details. Refer section 3.2 for more details.

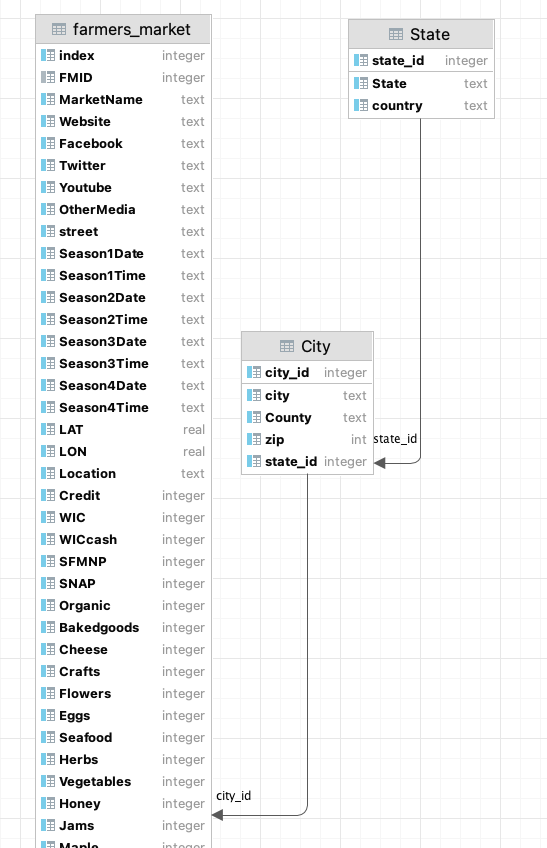
* **U2(which can not be achieved):** Identify records with high fidelity and faithfulness across demographics and firmographics?   
  Farmer’s data has columns which store social media details and other firmographics. However, many of the markets do not have their social website in the database as well as have missing information about goods they sell. As a result we have empty cells.

To solve this use case we would need to detect the outliers and we could use clustering or other machine learning techniques to do so. However, even if we went ahead and tried cleaning the data(url parsing etc.) it’s very hard to develop a technique as many primary data features are empty. Since filling empty cells is not possible with data cleaning, we conclude that data cleaning itself is not sufficient in all capacities to achieve our use case.

1. **Dataset Description**

**ER Diagram-**

We observed city information is duplicate across dataset so normalized it using RDBMS 3rd normal form to City and State tables.

****

**3.1 Schema**

DDL Statements-

**create table farmers\_market**

**(**

**"index" INTEGER,**

**FMID INTEGER**

**constraint farmers\_market\_pk**

**unique,**

**MarketName TEXT,**

**Website TEXT,**

**Facebook TEXT,**

**Twitter TEXT,**

**Youtube TEXT,**

**OtherMedia TEXT,**

**street TEXT,**

**Season1Date TEXT,**

**Season1Time TEXT,**

**Season2Date TEXT,**

**Season2Time TEXT,**

**Season3Date TEXT,**

**Season3Time TEXT,**

**Season4Date TEXT,**

**Season4Time TEXT,**

**LAT REAL,**

**LON REAL,**

**Location TEXT,**

**Credit INTEGER,**

**WIC INTEGER,**

**WICcash INTEGER,**

**SFMNP INTEGER,**

**SNAP INTEGER,**

**Organic INTEGER,**

**Bakedgoods INTEGER,**

**Cheese INTEGER,**

**Crafts INTEGER,**

**Flowers INTEGER,**

**Eggs INTEGER,**

**Seafood INTEGER,**

**Herbs INTEGER,**

**Vegetables INTEGER,**

**Honey INTEGER,**

**Jams INTEGER,**

**Maple INTEGER,**

**Meat INTEGER,**

**Nursery INTEGER,**

**Nuts INTEGER,**

**Plants INTEGER,**

**Poultry INTEGER,**

**Prepared INTEGER,**

**Soap INTEGER,**

**Wine INTEGER,**

**Coffee INTEGER,**

**Beans INTEGER,**

**Fruits INTEGER,**

**Grains INTEGER,**

**Juices INTEGER,**

**Mushrooms INTEGER,**

**PetFood INTEGER,**

**Tofu INTEGER,**

**WildHarvested INTEGER,**

**updateTime TIMESTAMP,**

**city\_id INTEGER default '' not null**

**);**

**create table City**

**(**

**city TEXT,**

**County TEXT,**

**zip INT,**

**city\_id integer**

**constraint City\_pk**

**primary key**

**constraint City\_farmer\_market\_city\_id\_fk**

**references farmers\_market (city\_id)**

**on update restrict on delete restrict,**

**state\_id integer**

**);**

**create table State**

**(**

**State TEXT,**

**country TEXT,**

**state\_id integer**

**constraint State\_pk**

**primary key**

**constraint State\_city\_\_fk**

**references City (state\_id)**

**on update restrict on delete restrict**

**);**

**create index ix\_farmers\_market\_index**

**on farmers\_market ("index");**

**3.2 DML statements:**

**CREATE INDEX ix\_farmers\_market\_index ON farmers\_market ("index");**

**CREATE TABLE City AS**

**SELECT DISTINCT city, State, County, zip**

**FROM farmers\_market;**

**DELETE from City WHERE city IS NULL OR city = '-';**

**DELETE from farmers\_market WHERE city IS NULL OR city = '-';**

***--32 rows affected in 8 ms***

**alter table CITY add int not null;**

**UPDATE city**

**SET city\_id = (**

**SELECT *row\_number*() over (partition by 1) FROM City cn**

**WHERE cn.city = city.city**

**);**

**ALTER TABLE farmers\_market ADD COLUMN city\_id INTEGER NOT NULL DEFAULT '';**

**UPDATE farmers\_market**

**SET city\_id = (**

**SELECT city\_id**

**FROM city**

**WHERE city.city = farmers\_market.city**

**);**

**CREATE TABLE State as select distinct State,'USA' as country, *row\_number*() over (partition by 1) as state\_id from City**

**UPDATE City**

**SET state\_id = (**

**SELECT state\_id**

**FROM State**

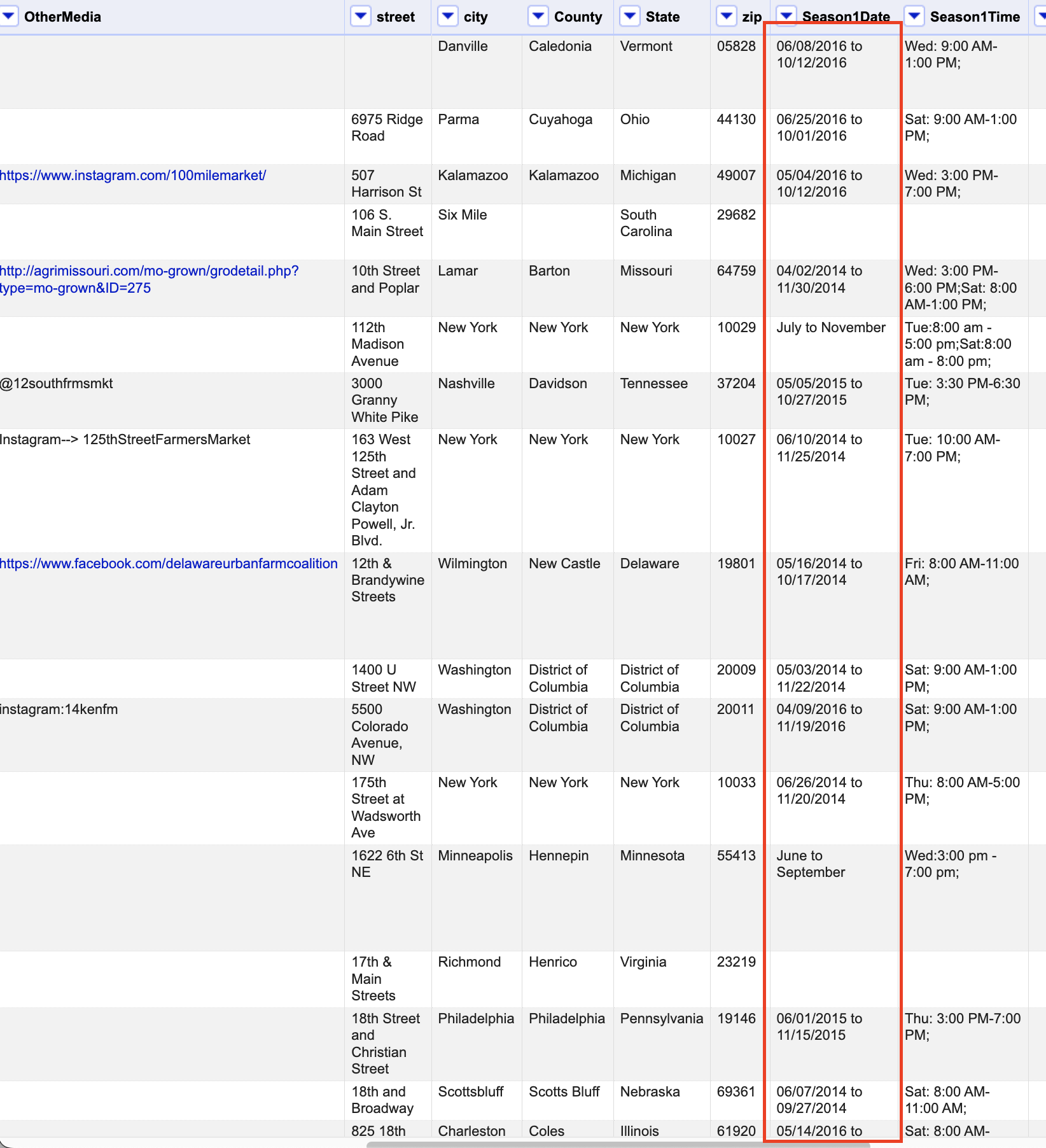
**WHERE city.State = State.State**

**);**

**ALTER TABLE City drop column state;**

For further details about the data initial cleanup using the pandas and open refine check [github](https://github.com/reethified/cs_513_project) repository.

1. **Data Quality Problems**
   1. As mentioned in U1.1 we are focussing on data issues in Seasonal Date and time columns. To give an example of one of the issues that we would be solving is:

****

As it is visible that there is “June to September” as well as other date formats present in the column which we would be cleaning to form a normalized time period column.

* 1. Similarly in “Season1Time” there is not any standard format which would require cleaning to process later.
  2. While plotting the heat map for U0 we came across outliers records based on latitude & longitude of each record.

Outliers are:

* 30 records have missing latitude & longitude values.
* Based on the [official coordinates](https://latitudelongitude.org/us/) system, the United States are in range: Latitude from 19.50139 to 64.85694 and longitude from -161.75583 to -68.01197. There are 53 records out of this range.
  1. Main table in the Farmer's Market Dataset is not in Normal Form. In order to remove redundant information(normalize) for U1.2 we need to extract city, state, postal code and country to a separate table. As a result of this activity we got to know that 32 farmer’s market have missing city and postal codes.
  2. Some other obvious data problems that we saw were in “Website” and “Facebook” columns where there were non-hyperlink text present.

1. **Project Implementation**
2. **Step 0:** Project logistics (repository, database setup) and communication preferences.
3. **Step 1:** Description of **Dataset(D)** and associated use case U0
4. **Step 2:** Profiling **D** to identify **Data Quality(DQ)** problems. We are using SQL and pandas to profile the data, assess the data and exploratory data analysis. We also use OpenRefine to manually profile the data. Steps taken for profiling:
   1. Identified data types of fields.
   2. Discover rows with missing values across the dataset.
   3. Plot Farmer’s Market distribution across the USA map using *latitude and longitude*.
   4. Identify various date formats in *seasonDates and times.*
   5. Identify rows with missing social media information (Youtube, Facebook, Twitter, Websites).
   6. Create database and tables using DDL commands in sqlite.
5. **Step 3:** Performing **Data Cleaning(DC)** proper: . Regex and Normalization techniques would be used.
6. **Step 4:** Checking that **D'** is "cleaner" than D. We would be running our U1 queries against the clean data **D’** to validate cleaning. Moreover We would come up with Integrity Constraints to ensure the columns have expected nature of value after cleaning.
7. **Tasks Assignment:** 
   1. Project’s logistical setup- Rahul
   2. High level analysis for finalizing dataset: Dinesh, Sahib, Rahul
   3. Database design, Table Schema, ER diagram- Rahul
   4. Use Case identification and prepare implementation plan- Dinesh, Rahul
   5. Exploratory data analysis and data quality checks- Dinesh, Rahul
   6. Phase-1 Report preparation- Dinesh, Rahul
   7. Use cases implementation:
      1. U0- Rahul
      2. U1.1- Dinesh
      3. U1.2- Rahul, Sahib
      4. U2  Dinesh, Sahib
   8. Final project report preparation- Dinesh, Sahib, Rahul