CS513 Final Project: Farmers Market Data Cleaning Workflow

Mark Berman, Sanjay Kumar, Jasdeep Duggal

Link to Clean data : https://uofi.box.com/s/rsukkuay8aadhxr4qlywp578pd6rtj2y

1) Introduction and Overview

In this report we use the basic data cleansing steps learnt from CS513 to present 2 use cases derived from the Farmer's Market dataset from the U.S. Department of Agriculture.

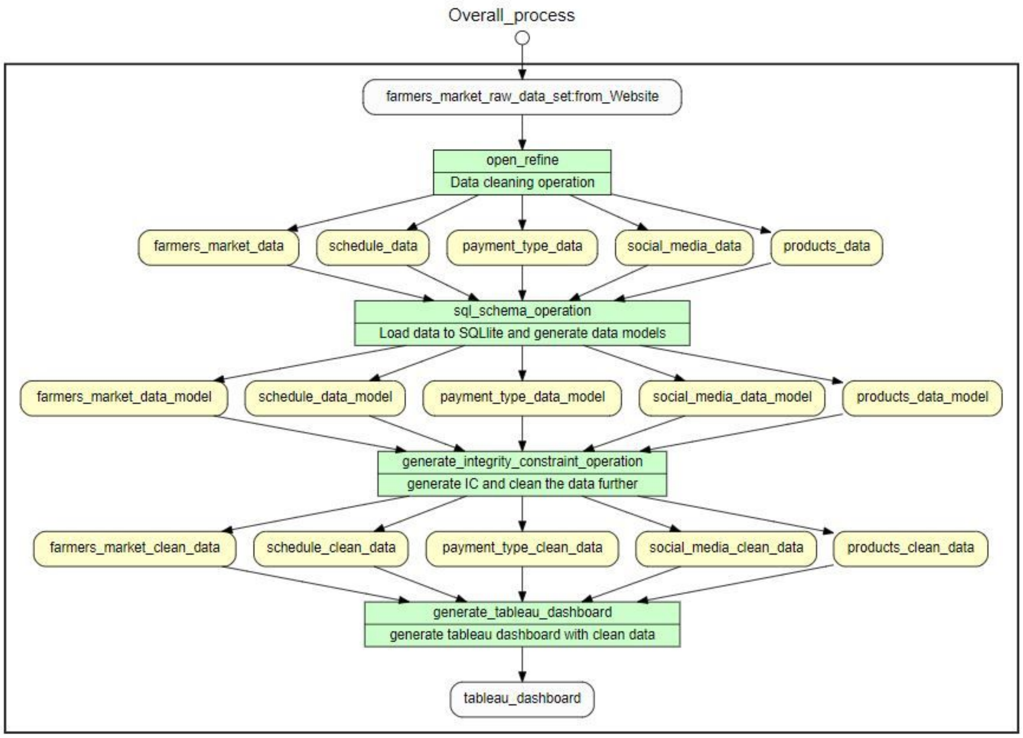
- Provide an intuitive way for customers to locate farmers markets based on product and payment type preferences.
- Understand how farmers markets are using social media to stimulate demand for their products

These use cases are presented as a set of interactive Tableau dashboards showcased at the end of the report.

The effectiveness of these dashboards depends an intuitive data model and data that is consistent and free from integrity constraint violations.

We decompose the single Farmers Market dataset into multiple subject areas that not only seem natural but also facilitate the creation of the Tableau dashboards. After decomposing the single dataset in to subject areas, we use Openrefine for column oriented data cleansing and transformations and we use SQL for integrity constraint discovery and remediation.

The following YesWorkflow diagram summarizes the data cleansing tasks we performed to produce the Tableau dashboards.

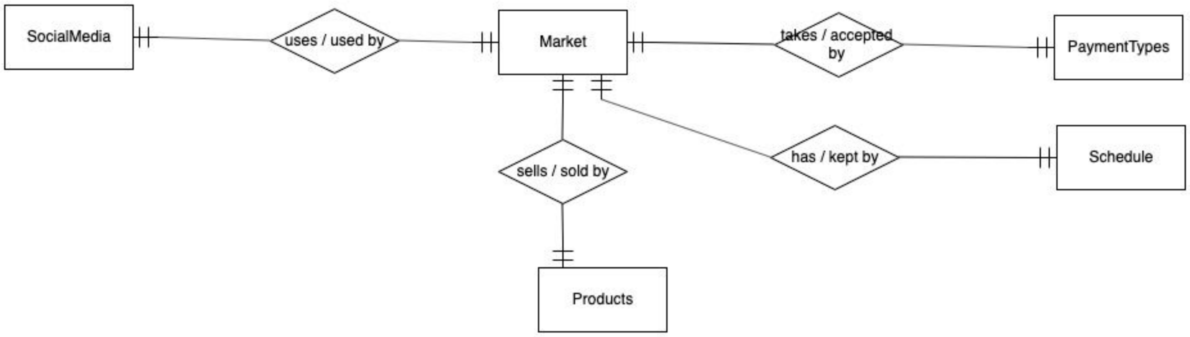


2) Initial Assessment of the dataset

The structure of Farmers Market dataset at first glance is trivial -- a single file with 59 columns. However, the dataset can be viewed as five distinct subject areas given the use cases described by this project.

1. Market - contains the each market’s name, address and geo-location.
2. Payment Types - contains the credit based payment types accepted by each market.
3. Products - contains the product types each market sells.
4. Schedule - contains the dates and times when each market is open for business.
5. Social Media - contains the URIs for each market’s social media presence.

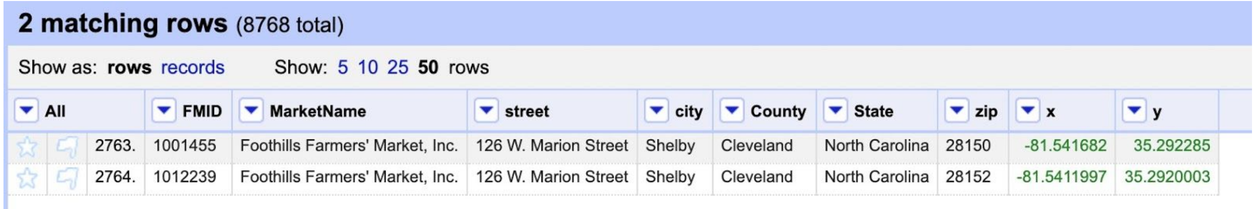
The ERD below shows the structure and relationship between the five subject areas.



Observed Data Quality Issues

Market Subject Area

- There are multiple instances where it appears that the same market appears in the dataset more than once. The following example illustrates this observation. The two market names are almost identical and their longitude and latitude are the same.

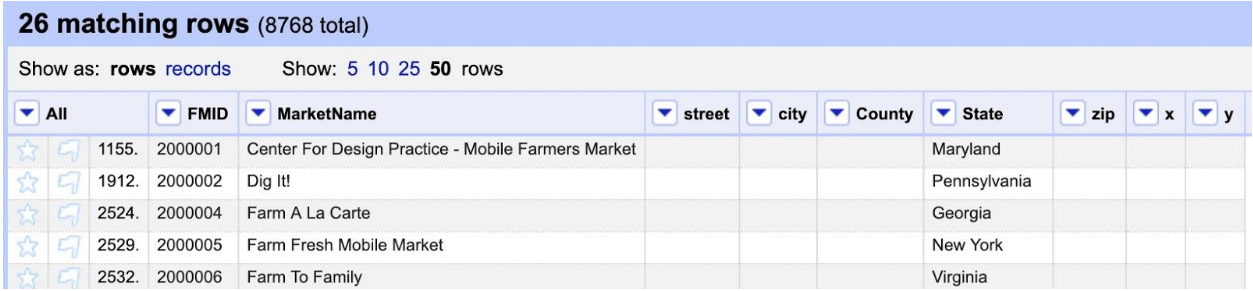


2 matching rows (8768 total)

Show as: **rows** records Show: 5 10 25 **50** rows

All	FMID	MarketName	street	city	County	State	zip	x	y
2763.	1001455	Foothills Farmers' Market, Inc.	126 W. Marion Street	Shelby	Cleveland	North Carolina	28150	-81.541682	35.292285
2764.	1012239	Foothills Farmers' Market, Inc.	126 W. Marion Street	Shelby	Cleveland	North Carolina	28152	-81.5411997	35.2920003

- There are multiple instances where geo-location values (e.g., street, city, longitude and latitude) are missing for a market. It is not possible to locate markets for a given geo-location without this data. The following example illustrates this observation.



26 matching rows (8768 total)

Show as: **rows** records Show: 5 10 25 **50** rows

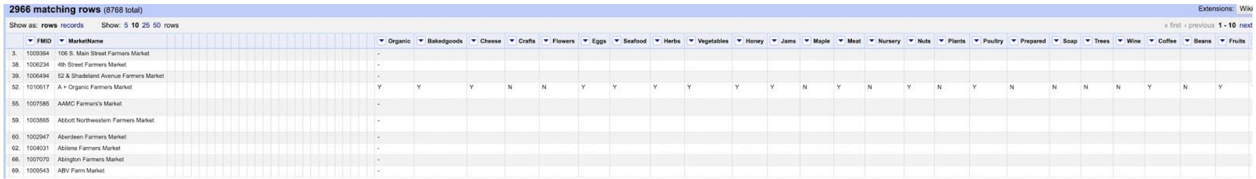
All	FMID	MarketName	street	city	County	State	zip	x	y
1155.	2000001	Center For Design Practice - Mobile Farmers Market				Maryland			
1912.	2000002	Dig It!				Pennsylvania			
2524.	2000004	Farm A La Carte				Georgia			
2529.	2000005	Farm Fresh Mobile Market				New York			
2532.	2000006	Farm To Family				Virginia			

Payment Types Subject Area

- No observed data quality issues

Products Subject Area

- There are multiple instances where products sold data is missing for a market. One cannot infer that a missing value for a product type means that type of product is not sold by that market. The following example illustrates this observation.



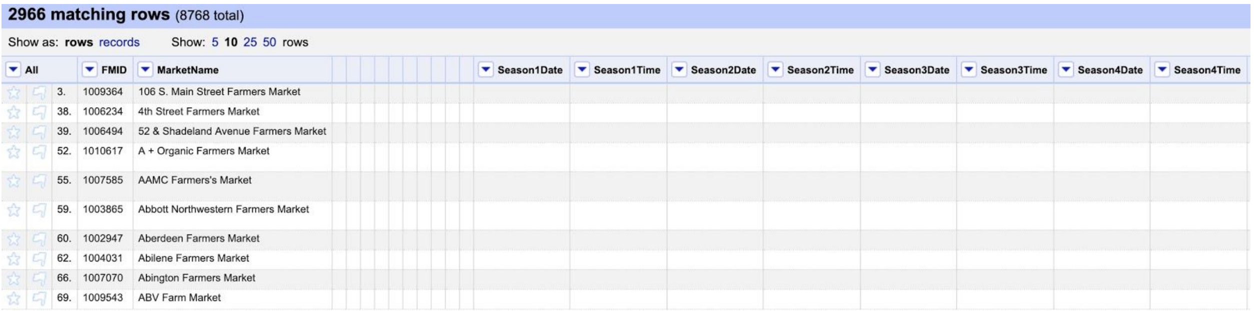
2966 matching rows (8768 total) Extensions: Wiki

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next

	FMID	MarketName	Organic	Bakedgoods	Cheese	Crafts	Flowers	Eggs	Seafood	Herbs	Vegetables	Honey	Jams	Maple	Meat	Nursery	Nuts	Plants	Poultry	Prepared	Soap	Trees	Wine	Coffee	Beans	Fruits
3.	1009364	106 S. Main Street Farmers Market	-																							
38.	1006234	4th Street Farmers Market	-																							
39.	1006494	52 & Shadeland Avenue Farmers Market	-																							
52.	1010617	A + Organic Farmers Market	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	N	Y	N	Y	N	Y	N	N	N	N	Y	N	Y
55.	1007585	AAMC Farmers's Market	-																							
59.	1003865	Abbott Northwestern Farmers Market	-																							
60.	1002947	Aberdeen Farmers Market	-																							
62.	1004031	Abilene Farmers Market	-																							
66.	1007070	Abington Farmers Market	-																							
69.	1009543	ABV Farm Market	-																							

Schedule Subject Area

- There are multiple instances where schedule data is missing for a market. Publishing data about a market without the market's date and times of availability is not very helpful to potential customers. The following example illustrates this point.



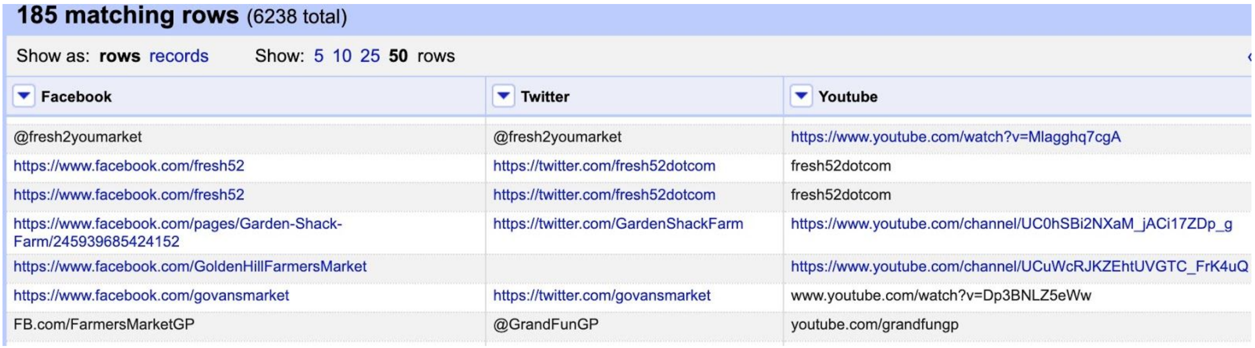
2966 matching rows (8768 total)

Show as: rows records Show: 5 10 25 50 rows

All		FMID	MarketName	Season1Date	Season1Time	Season2Date	Season2Time	Season3Date	Season3Time	Season4Date	Season4Time
☆	3.	1009364	106 S. Main Street Farmers Market								
☆	38.	1006234	4th Street Farmers Market								
☆	39.	1006494	52 & Shadeland Avenue Farmers Market								
☆	52.	1010617	A + Organic Farmers Market								
☆	55.	1007585	AAMC Farmers's Market								
☆	59.	1003865	Abbott Northwestern Farmers Market								
☆	60.	1002947	Aberdeen Farmers Market								
☆	62.	1004031	Abilene Farmers Market								
☆	66.	1007070	Abington Farmers Market								
☆	69.	1009543	ABV Farm Market								

Social Media Subject Area

- There are multiple instances where Facebook, Twitter and Youtube URIs are inconsistent or invalid. Inconsistency makes it more difficult to programmatically validate the correctness of URIs. Invalid URIs are not helpful to potential customers.



185 matching rows (6238 total)

Show as: **rows** records   Show: 5 10 25 **50** rows

Facebook	Twitter	Youtube
@fresh2youmarket	@fresh2youmarket	https://www.youtube.com/watch?v=Mlagghq7cgA
https://www.facebook.com/fresh52	https://twitter.com/fresh52dotcom	fresh52dotcom
https://www.facebook.com/fresh52	https://twitter.com/fresh52dotcom	fresh52dotcom
https://www.facebook.com/pages/Garden-Shack-Farm/245939685424152	https://twitter.com/GardenShackFarm	https://www.youtube.com/channel/UC0hSBi2NXaM_jACi17ZDp_g
https://www.facebook.com/GoldenHillFarmersMarket		https://www.youtube.com/channel/UCuWcRJKZEhtUVGTC_FrK4uQ
https://www.facebook.com/govansmarket	https://twitter.com/govansmarket	www.youtube.com/watch?v=Dp3BNLZ5eWw
FB.com/FarmersMarketGP	@GrandFunGP	youtube.com/grandfungp

- It should be noted the use cases for this project only depend on the presence or absence of a URI and not on URI consistency or correctness.

3) Data Cleaning methods and process

OpenRefine Based Data Cleansing

The overall data cleaning process is summarized in a YewWorkflow diagram at the following link.

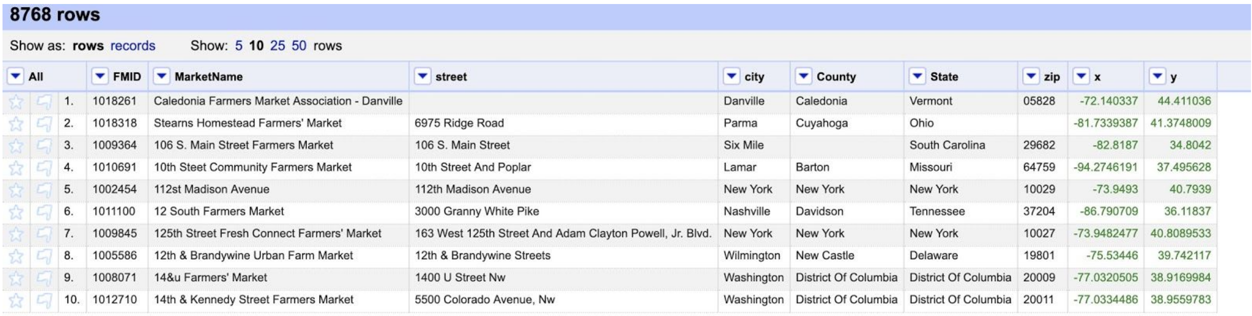
https://github.com/markcb2/cs513_datacleansing/blob/master/overall.JPG .

What follows is a summarization of the OpenRefine based data cleansing operations for each of the five subject areas.

Market Subject Area

Data Cleansing Operation	**Impacted Columns**
Remove columns	Website, Facebook, Twitter, YouTube, OtherMedia, Remove, Credit, WIC, WICcash, SFMNP, SNAP, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time, 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, WildHarvested, updateTime
Trim and collapse consecutive whitespace	FMID, MarketName, street, city, County, State, zip, x, y
Convert to Title Case	MarketName, street, city, County, State, zip
Convert to Number	x, y

The following screen capture shows the Market subject area after the above OpenRefine based data cleansing operations have been performed.



8768 rows

Show as: **rows** records   Show: 5 **10** 25 50 rows

All	FMID	MarketName	street	city	County	State	zip	x	y
1.	1018261	Caledonia Farmers Market Association - Danville		Danville	Caledonia	Vermont	05828	-72.140337	44.411036
2.	1018318	Stearns Homestead Farmers' Market	6975 Ridge Road	Parma	Cuyahoga	Ohio		-81.7339387	41.3748009
3.	1009364	106 S. Main Street Farmers Market	106 S. Main Street	Six Mile		South Carolina	29682	-82.8187	34.8042
4.	1010691	10th Steet Community Farmers Market	10th Street And Poplar	Lamar	Barton	Missouri	64759	-94.2746191	37.495628
5.	1002454	112st Madison Avenue	112th Madison Avenue	New York	New York	New York	10029	-73.9493	40.7939
6.	1011100	12 South Farmers Market	3000 Granny White Pike	Nashville	Davidson	Tennessee	37204	-86.790709	36.11837
7.	1009845	125th Street Fresh Connect Farmers' Market	163 West 125th Street And Adam Clayton Powell, Jr. Blvd.	New York	New York	New York	10027	-73.9482477	40.8089533
8.	1005586	12th & Brandywine Urban Farm Market	12th & Brandywine Streets	Wilmington	New Castle	Delaware	19801	-75.53446	39.742117
9.	1008071	14&u Farmers' Market	1400 U Street Nw	Washington	District Of Columbia	District Of Columbia	20009	-77.0320505	38.9169984
10.	1012710	14th & Kennedy Street Farmers Market	5500 Colorado Avenue, Nw	Washington	District Of Columbia	District Of Columbia	20011	-77.0334486	38.9559783

The generated post-refinement OpenRefine recipe for this subject area is viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/blob/master/farmers_market_base_table_history.txt

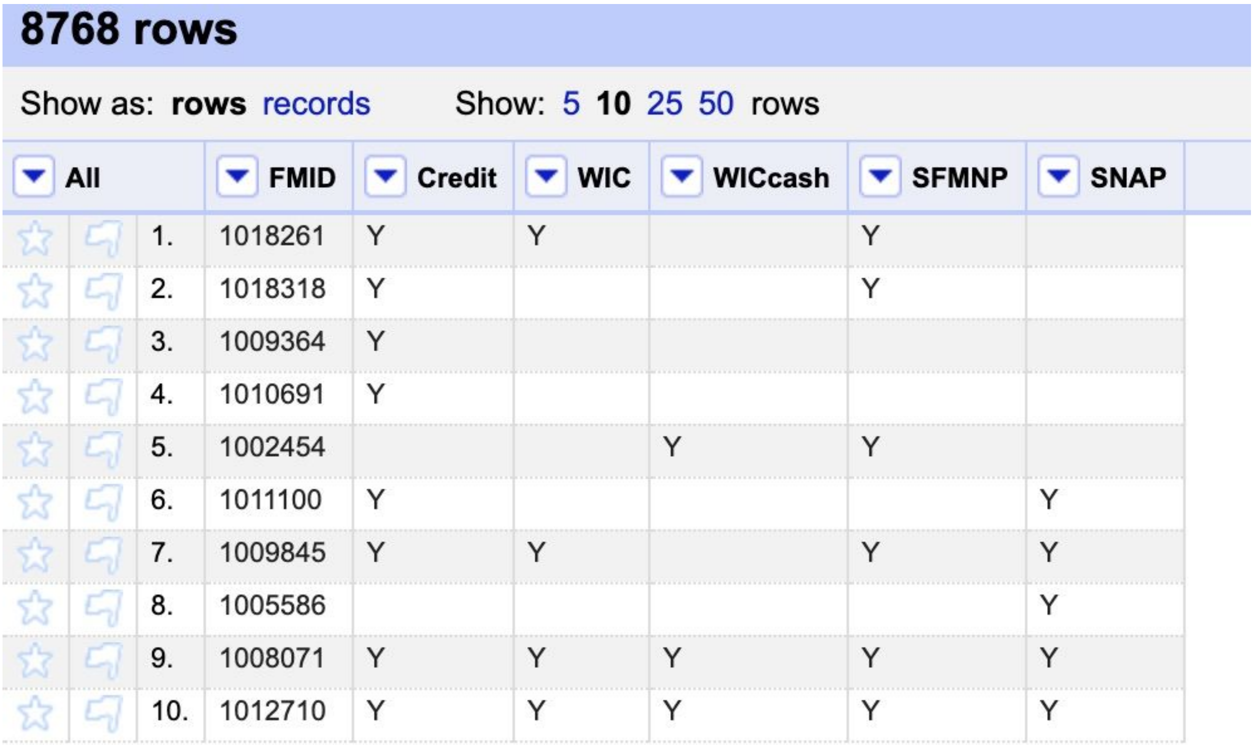
The YesWorkflow linear and parallel diagrams, and YW script that produced them for this subject area are viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/tree/master/yw_farmers_market_base_artifacts

Payment Types Subject Area

Data Cleansing Operation	**Impacted Columns**
Remove columns	MarketName, street, city, County, State, zip, x, y, Website, Facebook, Twitter, YouTube, OtherMedia, Remove, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time, 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, WildHarvested, updateTime
Trim and collapse consecutive whitespace	FMID, Credit, WIC, WICcash, SFMNP, SNAP
Convert to Upper Case	Credit, WIC, WICcash, SFMNP, SNAP
Replace N with empty string Done to facilitate calculation of count metrics for the Tableau visualizations.	Credit, WIC, WICcash, SFMNP, SNAP

The following screen capture shows the Payment Types subject area after the above OpenRefine based data cleansing operations have been performed.



All			FMID	Credit	WIC	WICcash	SFMNP	SNAP
☆		1.	1018261	Y	Y		Y	
☆		2.	1018318	Y			Y	
☆		3.	1009364	Y				
☆		4.	1010691	Y				
☆		5.	1002454			Y	Y	
☆		6.	1011100	Y				Y
☆		7.	1009845	Y	Y		Y	Y
☆		8.	1005586					Y
☆		9.	1008071	Y	Y	Y	Y	Y
☆		10.	1012710	Y	Y	Y	Y	Y

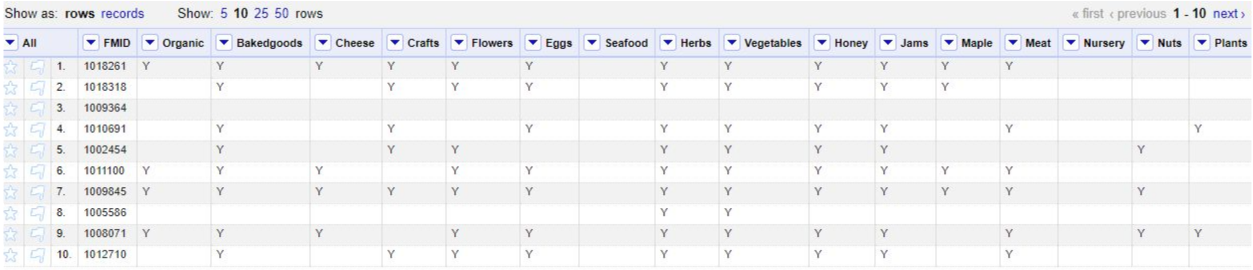
The generated post-refinement OpenRefine recipe for this subject area is viewable at the following link.
https://github.com/markcb2/cs513_datacleansing/blob/master/paymentType_history.txt

The YesWorkflow linear and parallel diagrams, and YW script that produced them for this subject area are viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/tree/master/yw_paymentType_artifacts

Products Subject Area

Cleaning of products type information from the dataset was straight forward. The objective of forming a data model for better data visualization was achieved by reordering and removal of the columns except for the products type column and farmers market unique identifier (FMID) column. Following is the screenshot of the products openrefine project after the data cleaning:



Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 10 next ›

All	FMID	Organic	Bakedgoods	Cheese	Crafts	Flowers	Eggs	Seafood	Herbs	Vegetables	Honey	Jams	Maple	Meat	Nursery	Nuts	Plants
1.	1018261	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y			
2.	1018318		Y		Y	Y	Y		Y	Y	Y	Y	Y				
3.	1009364																
4.	1010691		Y		Y		Y		Y	Y	Y	Y		Y			Y
5.	1002454		Y		Y	Y			Y	Y	Y	Y				Y	
6.	1011100	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y			
7.	1009845	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y		Y	
8.	1005586								Y	Y							
9.	1008071	Y	Y	Y		Y	Y		Y	Y	Y	Y		Y		Y	Y
10.	1012710		Y		Y	Y	Y		Y	Y	Y	Y		Y			

Data Cleansing Operation	**Impacted Columns**
Reorder Columns	FMID, 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, WildHarvested
Removed Columns	MarketName, Website, Facebook, Twitter, YouTube, OtherMedia, street, city, County, State, zip, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time, x, y, Location, SFMNP
Trim and collapse consecutive whitespace	All Columns

The generated post-refinement OpenRefine recipe for this subject area is viewable at the following link.

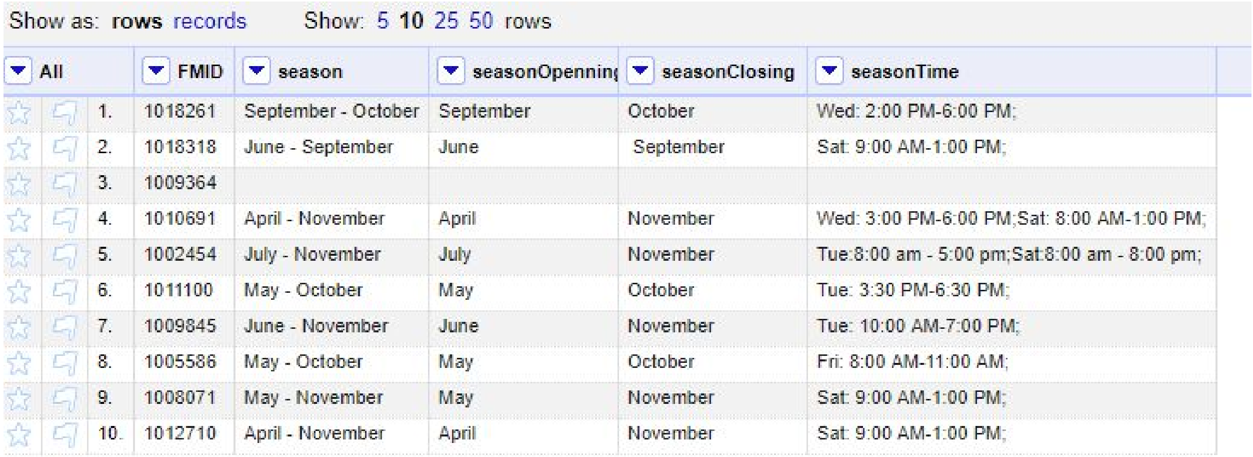
https://github.com/markcb2/cs513_datacleansing/blob/master/productType_history.txt .

The YesWorkflow linear and parallel diagrams, and YW script that produced them for this subject area are viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/tree/master/yw_product_artifacts

Schedule Subject Area

Refining and extracting schedule information for each farmers market was challenging with open refine. After a series of reordering and extraction of opening and closing seasons, we were able to provide meaningful data to create a data model in SQL.
Following is the screenshot of the schedule openrefine project after the data cleaning:


Show as: **rows** records Show: 5 **10** 25 50 rows

		All	FMID	season	seasonOpenning	seasonClosing	seasonTime
☆		1.	1018261	September - October	September	October	Wed: 2:00 PM-6:00 PM;
☆		2.	1018318	June - September	June	September	Sat: 9:00 AM-1:00 PM;
☆		3.	1009364				
☆		4.	1010691	April - November	April	November	Wed: 3:00 PM-6:00 PM;Sat: 8:00 AM-1:00 PM;
☆		5.	1002454	July - November	July	November	Tue:8:00 am - 5:00 pm;Sat:8:00 am - 8:00 pm;
☆		6.	1011100	May - October	May	October	Tue: 3:30 PM-6:30 PM;
☆		7.	1009845	June - November	June	November	Tue: 10:00 AM-7:00 PM;
☆		8.	1005586	May - October	May	October	Fri: 8:00 AM-11:00 AM;
☆		9.	1008071	May - November	May	November	Sat: 9:00 AM-1:00 PM;
☆		10.	1012710	April - November	April	November	Sat: 9:00 AM-1:00 PM;

Data Cleansing Operation	**Impacted Columns**
Reorder Columns	FMID,
Removed columns	MarketName, Website, Facebook, Twitter, YouTube, OtherMedia, street, city, County, State, zip, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time, x, y, Location, SFMNP, All product types
New Columns	Season, seasonOpenning, seasonClosing, SeasonTime
Trim and collapse consecutive whitespace	All columns

The generated post-refinement OpenRefine recipe for this subject area is viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/blob/master/schedule_history.txt

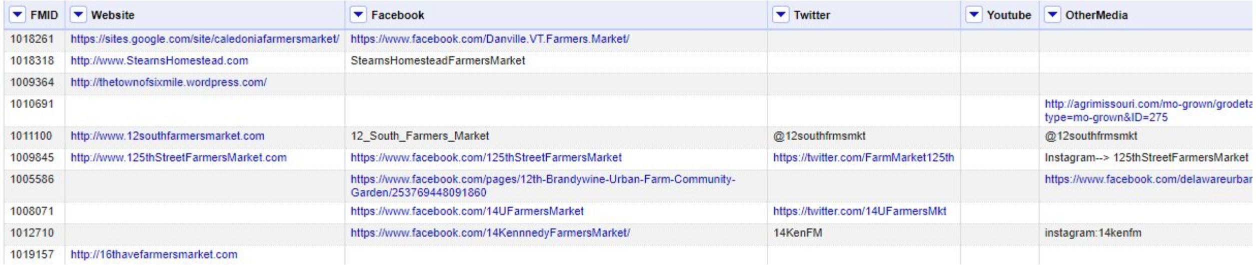
The YesWorkflow linear and parallel diagrams, and YW script that produced them for this subject area are viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/tree/master/yw_schedule_artifacts

Social Media Subject Area

Cleaning of social media information from the initial dataset was straight forward too. The objective of forming a data model for better data visualization was achieved by reordering and removal of the columns except for the columns containing social media information and farmers market unique identifier (FMID) column. As part of the cleaning, rows with no social media data were removed. Post refinement, it was found that there are **6238** farmers market which has a social media presence across USA.

Following is the screenshot of the social media openrefine project after the data cleaning:



FMID	Website	Facebook	Twitter	Youtube	OtherMedia
1018261	https://sites.google.com/site/caledoniafarmersmarket/	https://www.facebook.com/Danville.VT.Farmers.Market/			
1018318	http://www.SteamsHomestead.com	SteamsHomesteadFarmersMarket			
1009364	http://thetownofsixmile.wordpress.com/				
1010691					http://agrimissouri.com/mo-grown/grodeta type=mo-grown&ID=275
1011100	http://www.12southfarmersmarket.com	12_South_Farmers_Market	@12southfrmsmkt		@12southfrmsmkt
1009845	http://www.125thStreetFarmersMarket.com	https://www.facebook.com/125thStreetFarmersMarket	https://twitter.com/FarmMarket125th		Instagram--> 125thStreetFarmersMarket
1005586		https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860			https://www.facebook.com/delawareurbar
1008071		https://www.facebook.com/14UFarmersMarket	https://twitter.com/14UFarmersMkt		
1012710		https://www.facebook.com/14KennnedyFarmersMarket/	14KenFM		instagram:14kenfm
1019157	http://16thavefarmersmarket.com				

Data Cleansing Operation	Impacted Columns
Reorder Columns	FMID,Website, Facebook, Twitter, YouTube, OtherMedia
Removed columns	MarketName, street, city, County, State, zip, Season1Date, Season1Time, Season2Date, Season2Time, Season3Date, Season3Time, Season4Date, Season4Time, x, y, Location, SFMNP, All product types
Trim and collapse consecutive whitespace	All columns

The generated post-refinement OpenRefine recipe for this subject area is viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/blob/master/socialMedia_history.txt

The YesWorkflow linear and parallel diagrams, and YW script that produced them for this subject area are viewable at the following link.

https://github.com/markcb2/cs513_datacleansing/tree/master/yw_socialMedia_artifacts

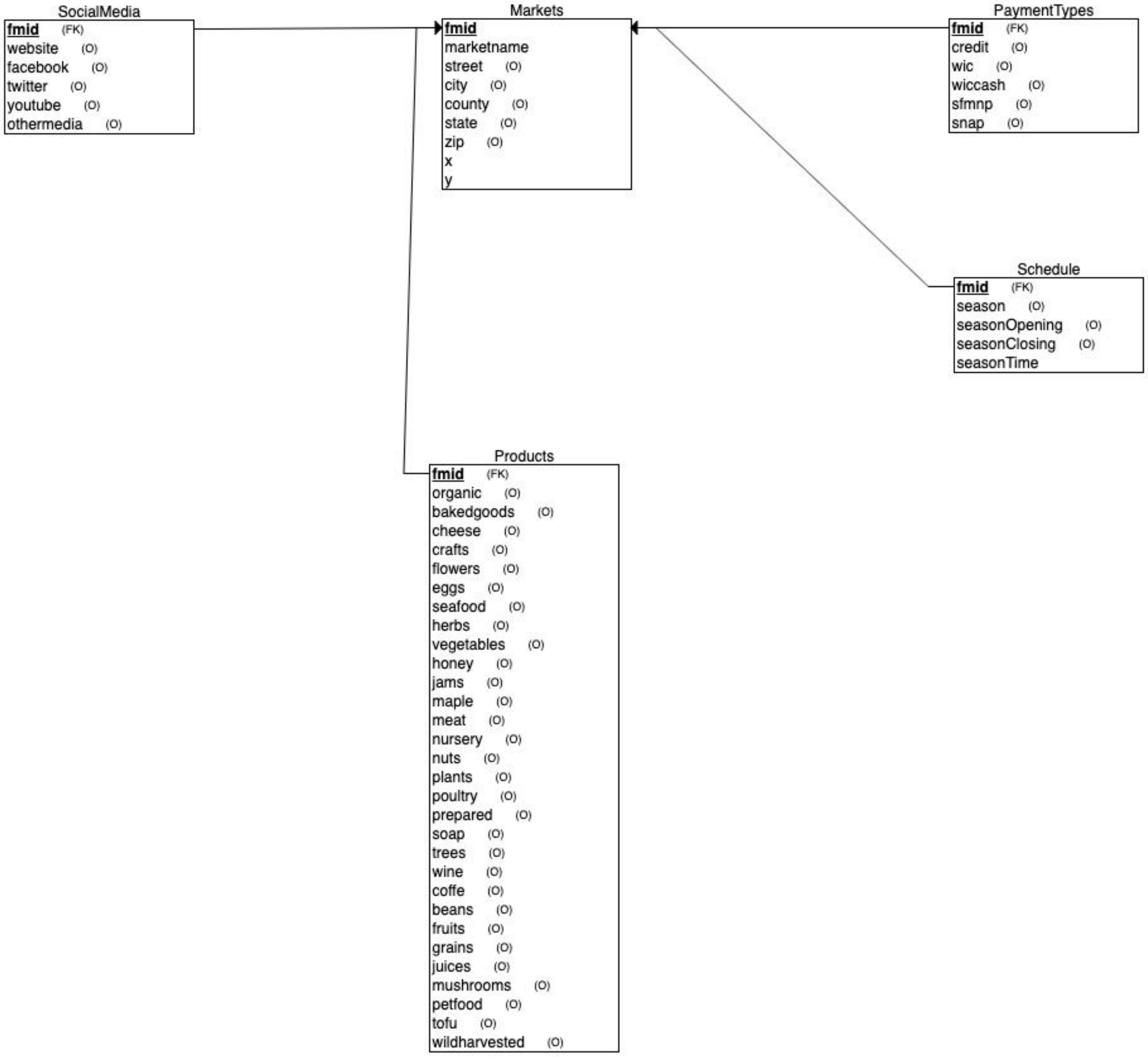
OpenRefine Data Cleaning Limitations and Alternative Approaches

OpenRefine is best suited for “column at a time” data cleansing. OpenRefine is not ideal for finding duplicate records or cases where multiple columns in a record need to be evaluated. Seventeen different integrity constraint checks were performed using SQL queries against a database schema implemented using SQLite. There were seven integrity constraint violations that were identified. These integrity constraint violations were remediated using SQL ***“Create Table <TableName> as select”*** statements. The integrity constraint checks and the remediation steps are described in the **Data Cleaning Results** section.

4) Data Cleaning Results

Relational Schema and Integrity Constraint Remediation.

OpenRefine was used to generate the SQL schema. The SQL schema along with its cleansed data was exported from OpenRefine and loaded into SQLite. A relational data diagram of the schema and the SQL commands used to generate the schema are shown below.



```
DROP TABLE IF EXISTS markets;
CREATE TABLE markets (
FMID VARCHAR(10) NOT NULL,
MarketName VARCHAR(100) NOT NULL,
street VARCHAR(100) NULL,
city VARCHAR(50) NULL,
County VARCHAR(50) NULL,
State VARCHAR(50) NULL,
zip VARCHAR(10) NULL,
x NUMERIC(12) NULL,
y NUMERIC(12) NULL
);
```

```
DROP TABLE IF EXISTS paymentTypes;
CREATE TABLE paymentTypes (
FMID VARCHAR(10) NOT NULL,
Credit VARCHAR(1) NULL,
WIC VARCHAR(1) NULL,
WICcash VARCHAR(1) NULL,
SFMNP VARCHAR(1) NULL,
SNAP VARCHAR(1) NULL
);
```

DROP TABLE IF EXISTS products; CREATE TABLE products (FMID VARCHAR(10) NOT NULL, Organic VARCHAR(1) NULL, Bakedgoods VARCHAR(1) NULL, Cheese VARCHAR(1) NULL, Crafts VARCHAR(1) NULL, Flowers VARCHAR(1) NULL, Eggs VARCHAR(1) NULL, Seafood VARCHAR(1) NULL, Herbs VARCHAR(1) NULL, Vegetables VARCHAR(1) NULL, Honey VARCHAR(1) NULL, Jams VARCHAR(1) NULL, Maple VARCHAR(1) NULL, Meat VARCHAR(1) NULL, Nursery VARCHAR(1) NULL, Nuts VARCHAR(1) NULL, Plants VARCHAR(1) NULL, Poultry VARCHAR(1) NULL, Prepared VARCHAR(1) NULL, Soap VARCHAR(1) NULL, Trees VARCHAR(1) NULL, Wine VARCHAR(1) NULL, Coffee VARCHAR(1) NULL, Beans VARCHAR(1) NULL, Fruits VARCHAR(1) NULL, Grains VARCHAR(1) NULL, Juices VARCHAR(1) NULL, Mushrooms VARCHAR(1) NULL, PetFood VARCHAR(1) NULL, Tofu VARCHAR(1) NULL, WildHarvested VARCHAR(1) NULL);	DROP TABLE IF EXISTS socialMedia; CREATE TABLE socialMedia (FMID VARCHAR(10) NOT NULL, Website VARCHAR(256) NULL, Facebook VARCHAR(256) NULL, Twitter VARCHAR(256) NULL, Youtube VARCHAR(256) NULL, OtherMedia VARCHAR(256) NULL);
DROP TABLE IF EXISTS schedule; CREATE TABLE schedule (FMID VARCHAR(10) NOT NULL, season VARCHAR(50) NULL, seasonOpenning VARCHAR(50) NULL, seasonClosing VARCHAR(50) NULL, seasonTime VARCHAR(100) NULL);	

The following table lists the seventeen integrity constraint checks that were performed and the seven places where there were violations. The actual integrity constraint SQL queries and the result sets returned by these queries are available at the following link.

https://github.com/markcb2/cs513_datacleansing/blob/master/ic_queries.sql

Integrity Constraint Check	**Violation Occurrences**
IC1: Markets records where there are at least two rows having the same ID, but different column values.	No violations.
IC2: Missing market names in the Markets table.	No violations.
IC3: Duplicate markets in the Markets table.	11 violations.
IC4: Records with Invalid US longitude or latitude in the Markets table. Cannot execute the geo-location use cases on those records that violate this constraint.	No violation.
IC5: Records with missing longitude or latitude in the Markets table. Cannot execute the geo-location use cases on those records that violate this constraint.	28 violations.
IC6: Social Media records where there are at least two rows having the same ID, but different column values.	No violations.
IC7: Invalid websites in the Social Media table. Valid web sites must have at least one character between "http(s)://" and the "." and at least two characters after the dot.	No violations.
IC8: Payment Type records where there are at least two rows having the same ID but different column values.	No violations.
IC9: Records with invalid payment type indicator values (valid values are 'Y' or null) in the Payment Types table. (‘N’ values are converted to empty strings as part of OpenRefine based data cleansing.)	No violations.
IC10: Schedule records where there are at	No violations.

least two rows having the same ID, but different column values.	
IC11: Schedule records where there is no or incomplete schedule information. Cannot execute the use case that displays schedule information for those records that violate this constraint.	3205 violations
IC12: Product records where there are at least two rows having the same ID, but different column values.	No violations.
IC13: Product records with no product information.	No violations.
IC14: Social Media records where its foreign key not found in Markets table after the Markets table was purged of integrity constraints.	37 violations.
IC15: Payment Types records where their foreign keys are not found in the Market table after the Markets table was purged of integrity constraints	39 violations.
IC16: Product records where their foreign keys are not found in the Market tables after the Markets table was purged of integrity constraints.	39 violations.
IC17: Schedule records where their foreign keys are not found in the Markets table after the Markets table was purged of integrity constraints.	39 violations.

The following table shows the record count for all five subject areas before and after integrity constraint remediation.

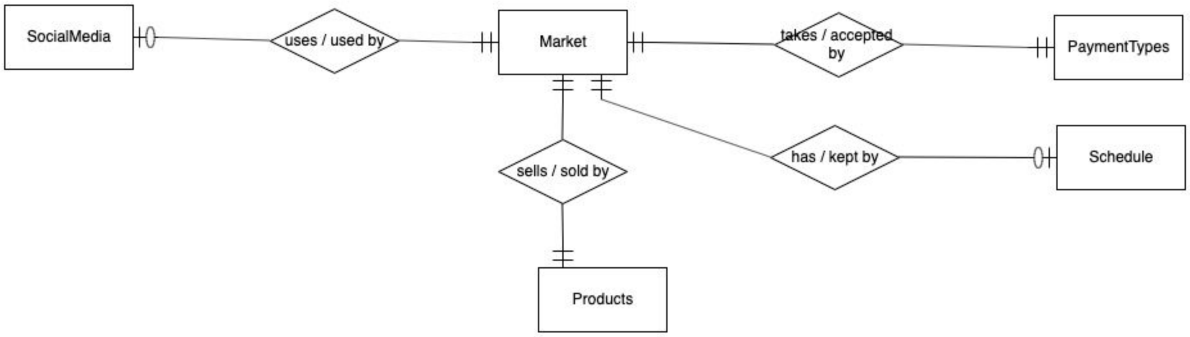
	Records Count Before IC Remediation	Record Count After IC Remediation
Markets	8768	8729
Payment Types	8768	8729

Products	8768	8729
Schedule	8768	5551
Social Media	6238	6201

The **“Create table <TableName> as select”** SQL statements used to remediate the integrity constraint violations are available at the following link.

https://github.com/markcb2/cs513_datacleansing/blob/master/ic_queries.sql

The combination of OpenRefine based data cleansing and integrity constraint remediation changes the cardinality of from 1:1 to 0:1 on two of the relationships of the subject area ERD. The following diagram illustrates the cardinality changes of the Markets to Schedule relationship and the Markets to Social Media relationship.

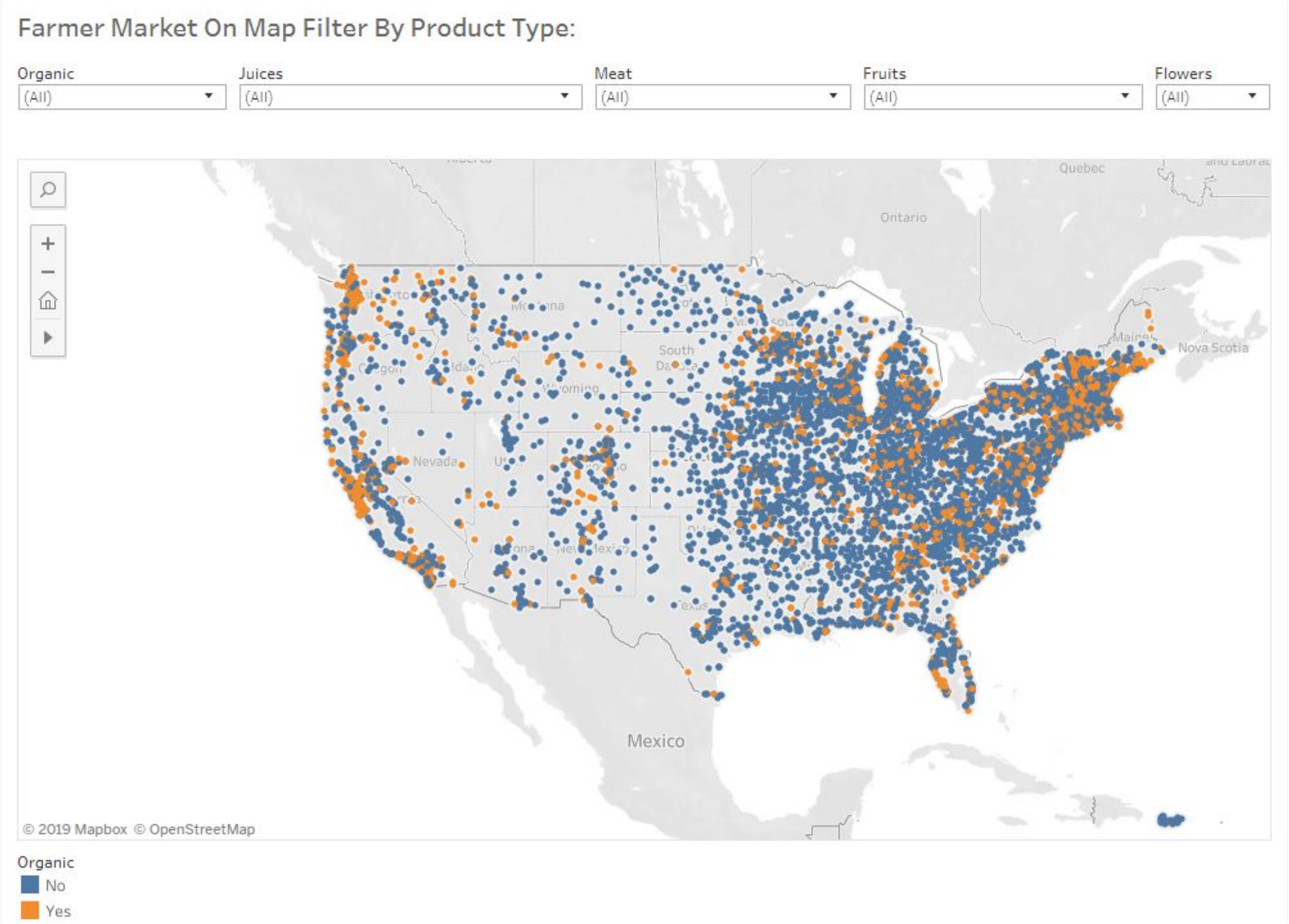


Use Cases

We choose tableau to represent the cleaned farmers market data on to below dashboards. It helped us in simplifying refined cleaned data into the very easily understandable format.

https://public.tableau.com/profile/jack.smith8848#!/vizhome/FarmerMarketDashbord/FarmerMarketOnMapFilterByProductType

https://public.tableau.com/profile/sanjay5224#!/vizhome/Dashboard-MapofFarmersMarketwithProductInfo/Dashboard

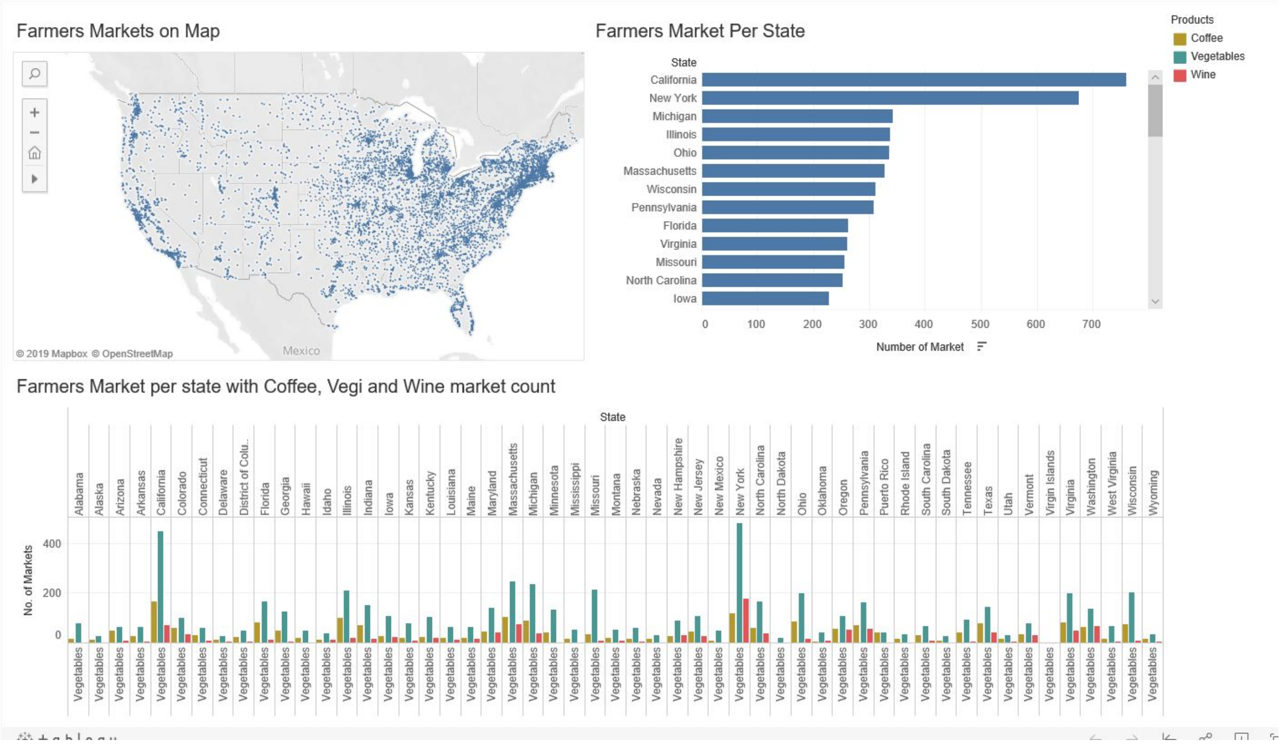


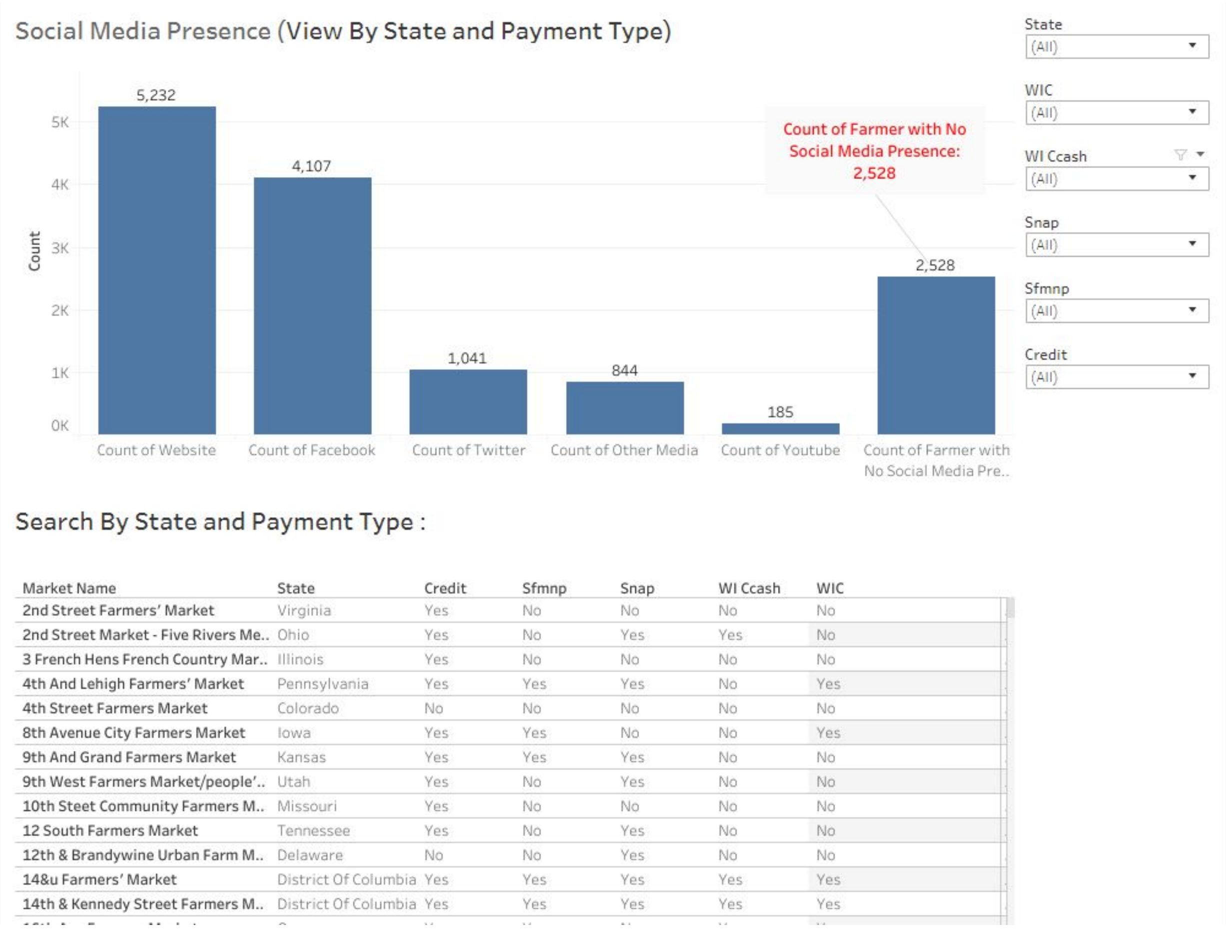
Farmer Market On Map Filter By Product Type:

Organic (All)
Juices (All)
Meat (All)
Fruits (All)
Flowers (All)

Organic
🟦 No
🟧 Yes

User Case: Find geolocation of farmer Market . Color coded Organic vs non Organic farmer. Search by different product type.


Farmers Markets on Map
© 2019 Mapbox © OpenStreetMap
Mexico
Farmers Market Per State
State
California
New York
Michigan
Illinois
Ohio
Massachusetts
Wisconsin
Pennsylvania
Florida
Virginia
Missouri
North Carolina
Iowa
0
100
200
300
400
500
600
700
Number of Market
Products
🟨 Coffee
🟩 Vegetables
🟥 Wine
Farmers Market per state with Coffee, Vegi and Wine market count
State
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Colu..
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virgin Islands
Virginia
Washington
West Virginia
Wisconsin
Wyoming
No. of Markets
400
200
0
Vegetables
tableau



Social Media Presence (View By State and Payment Type)

State: (All)

WIC: (All)

WI Ccash: (All)

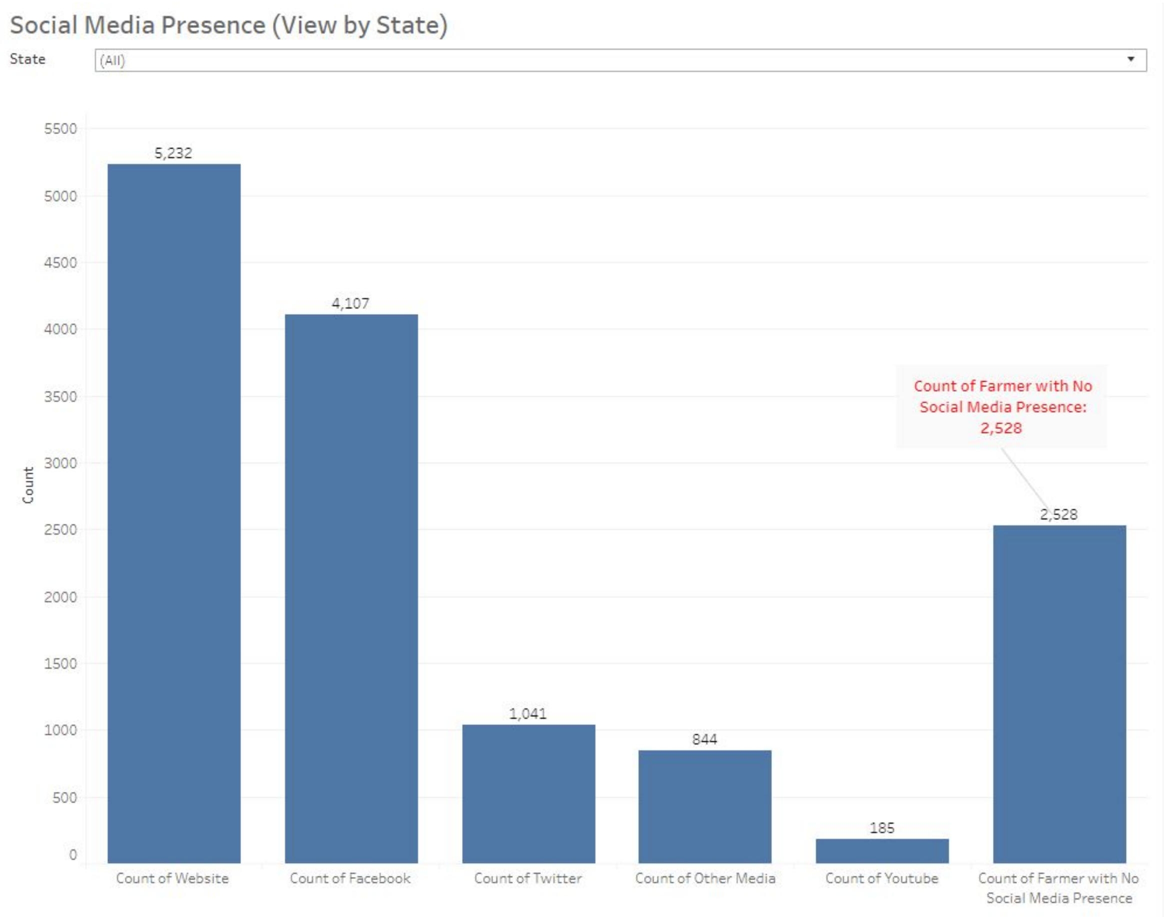
Snap: (All)

Sfmnp: (All)

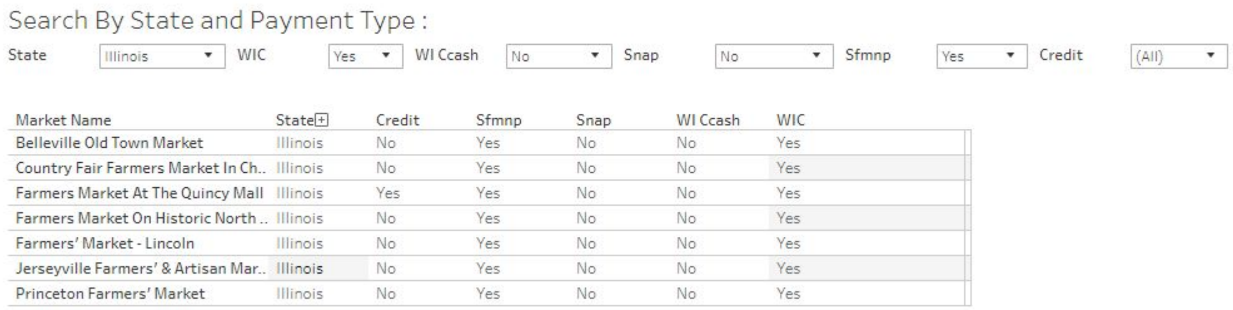
Credit: (All)

Search By State and Payment Type :

Market Name	State	Credit	Sfmnp	Snap	WI Ccash	WIC
2nd Street Farmers' Market	Virginia	Yes	No	No	No	No
2nd Street Market - Five Rivers Me..	Ohio	Yes	No	Yes	Yes	No
3 French Hens French Country Mar..	Illinois	Yes	No	No	No	No
4th And Lehigh Farmers' Market	Pennsylvania	Yes	Yes	Yes	No	Yes
4th Street Farmers Market	Colorado	No	No	No	No	No
8th Avenue City Farmers Market	Iowa	Yes	Yes	No	No	Yes
9th And Grand Farmers Market	Kansas	Yes	Yes	Yes	No	No
9th West Farmers Market/people'..	Utah	Yes	No	Yes	No	No
10th Steet Community Farmers M..	Missouri	Yes	No	No	No	No
12 South Farmers Market	Tennessee	Yes	No	Yes	No	No
12th & Brandywine Urban Farm M..	Delaware	No	No	Yes	No	No
14&u Farmers' Market	District Of Columbia	Yes	Yes	Yes	Yes	Yes
14th & Kennedy Street Farmers M..	District Of Columbia	Yes	Yes	Yes	Yes	Yes



Count of Social Media Presence . User can search by State.



Search By State and Payment Type :

State Illinois ▾ WIC Yes ▾ WI Ccash No ▾ Snap No ▾ Sfmnp Yes ▾ Credit (All) ▾

Market Name	State	Credit	Sfmnp	Snap	WI Ccash	WIC
Belleville Old Town Market	Illinois	No	Yes	No	No	Yes
Country Fair Farmers Market In Ch..	Illinois	No	Yes	No	No	Yes
Farmers Market At The Quincy Mall	Illinois	Yes	Yes	No	No	Yes
Farmers Market On Historic North ..	Illinois	No	Yes	No	No	Yes
Farmers' Market - Lincoln	Illinois	No	Yes	No	No	Yes
Jerseyville Farmers' & Artisan Mar..	Illinois	No	Yes	No	No	Yes
Princeton Farmers' Market	Illinois	No	Yes	No	No	Yes

Market Directory where user can search by state and payment type.
This will help user to locate nearest farmer's market place.

5) Conclusions

Data cleaning can be performed on a data set using open source tools such as OpenRefine and ubiquitous tool such as SQL in a sort span of time. OpenRefine is simple and quick to learn. In order to clean a raw set of data, a combination of various open source tools such as OpenRefine, SQL and OR2YW, are used, which is great but can't be used in industry for professional use cases especially on large and more complex ETL pipeline. Using OR2YW is an interesting concept but its not practical for large data cleaning pipelines such as ETL, as the drawn YesWorkflow would not be intuitive and doesn't give an overall picture.

After data cleaning was performed, the cleansed data was loaded on to tableau to provide a fair representation of each farmers market location on the USA map. It eases the viewer to locate a farmers market and draw a quick conclusion on the types of products sold and various payment types used.

The dashboard also provided various other conclusion such as there are over 2000 farmers market across the USA, which doesn't have a social media presence and are not able to benefit from digital marketing. It would be a huge socio-economic opportunity for farmers and digital marketers to expand their business on.