CS 513 Theory & Practice of Data Cleaning

Final Project: Data Cleaning Project

Author: Varun Sharma

Email :Varun4@illinois.edu

Table of Contents

1. Introduction

In this report farmers market data is analyzed and cleansed to solve some common use cases related to farmer market information.

This report is done as part of Individual submission of data cleaning project done by varun sharma at UIUC.

The tools used in the project are OpenRefine for clustering and refining data, SQLite for relational database, Trifacta Data Wrangler and YesWorkflow for provenance.

2. Overview and Initial Assessment

i. *Dataset*

I am using data from the United States Department of Agriculture. 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.

URL for the data set:

https://www.ams.usda.gov/local-food-directories/farmersmarkets

ii. *Proposed Use Cases*

Proposed use case for this dataset is to provide easy access to consumers to clean Farmer market data where they can easily browse trusted data to figure out details of food markets in the area. Consumers can have more details about the farmer markets.

This data can be fed to other web application or IA engines/automated assistant like Alexa or google voice to make them smarter. This use case is not part of the project but worth mentioning because questions like which markets have organic food zip? 2. Which market are open during night time of 8-11pm? can be answered easily using the clean data set used in this project.

iii. *Data Cleaning Goals*

The main goal for this project is to clean data based on data quality dimensions using farmer's market dataset.

I. **Completeness**: There are lot of columns with missing information. For example youtube column has 8558 null values which lowers the completeness score for the data.

II. **Consistency:** Data set has consistency issues like values NA,N.A,_ are used for missing values. Also, city values like ST. Louis and ST Louis exist in the dataset.

III. **Uniqueness**: Data set scores high on unique as there are no duplicates in the data set. Though there are multiple records for same store but their location are different in country.

IV. **Validity:** There are some validity problems in data set like Facebook column has URL and some places it doesn't have typical URL format.

V. **Accuracy:** Accuracy % drops because of inconsistent formats used for the fields for example:

VI. Update time field =201 has only year mentioned in it in one record and most of records have

VII. **Timeliness:** Data is downloaded from the website directly and cleansed which has some delay but it's up to date as its maintained by the United States Department of Agriculture.

Following whitepaper was consulted for this: https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

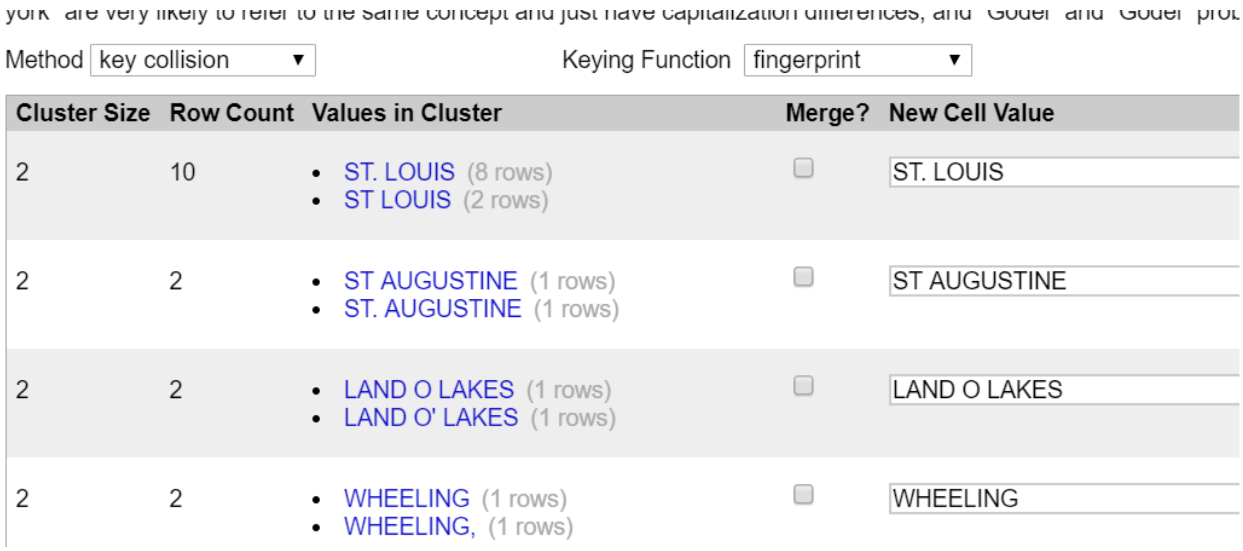
iv. High Level Data Quality Issues:

As discussed above data set has multiple issues. Following are the major high-level issues in the data set:

I. Lots of leading and trailing whitespaces in column values.
II. Markets with missing information like youtube
III. Standardization for date/time formats.
IV. Standard values for city names in the data set.

v. Data Cleaning Tools and Approach

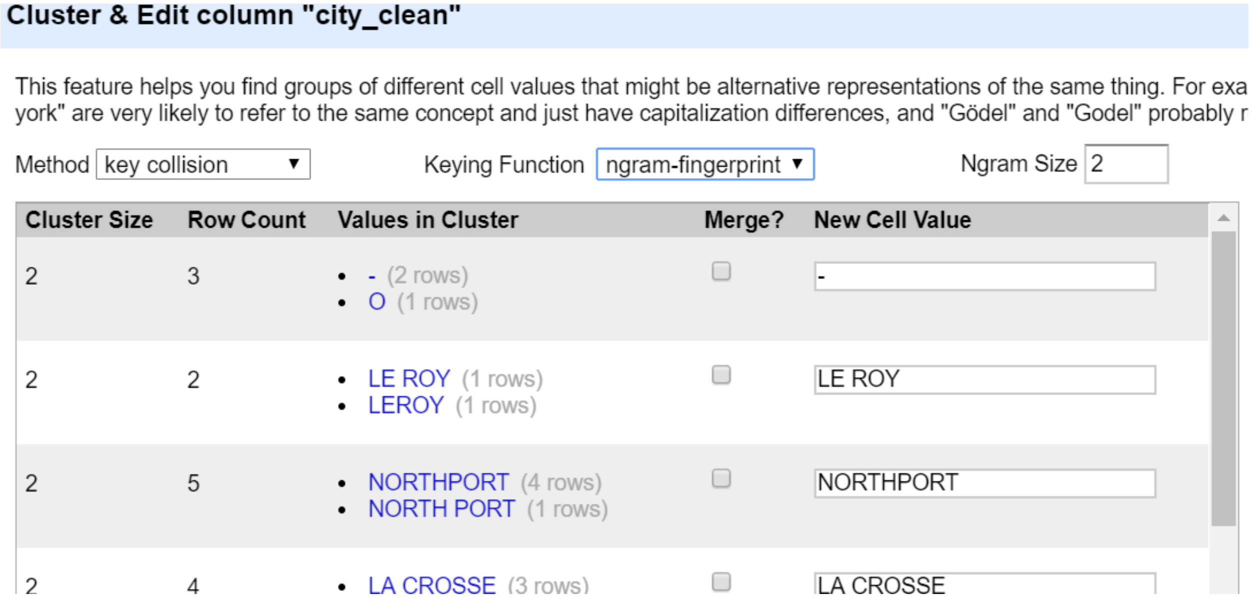
OpenRefine is used in this project for common data cleaning steps like removing white spaces, line breaks and doing basic transformations. Clustering capability helped to achieve data consistency goal for data quality. Clustering in openrefine was used for columns like county, cities:



york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" prob

Method key collision ▼ Keying Function fingerprint ▼

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	10	• ST. LOUIS (8 rows) • ST LOUIS (2 rows)	□	ST. LOUIS
2	2	• ST AUGUSTINE (1 rows) • ST. AUGUSTINE (1 rows)	□	ST AUGUSTINE
2	2	• LAND O LAKES (1 rows) • LAND O' LAKES (1 rows)	□	LAND O LAKES
2	2	• WHEELING (1 rows) • WHEELING, (1 rows)	□	WHEELING



Cluster & Edit column "city_clean"

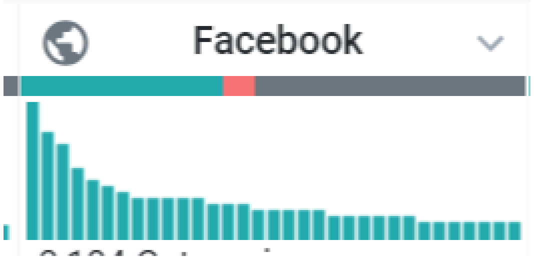
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For exa york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably r

Method key collision ▼ Keying Function ngram-fingerprint ▼ Ngram Size 2

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	3	• - (2 rows) • O (1 rows)	□	-
2	2	• LE ROY (1 rows) • LEROY (1 rows)	□	LE ROY
2	5	• NORTHPORT (4 rows) • NORTH PORT (1 rows)	□	NORTHPORT
2	4	• LA CROSSE (3 rows)	□	LA CROSSE

After OpenRefine step Trifecta Data Wrangler is used to check data quality. I felt trifacta data wrangler's visual representation is better than open refine. Wrangler gives capability to profile data column wise and understand data better.

Each column data pattern can be seen in the UI



Red here marks invalid values or mismatched values.

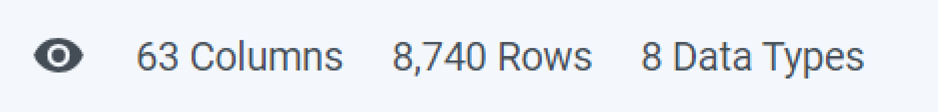
I have used YesWorkflow to create a graphical representation of the data cleaning process of the farmers market data. With http://try.yesworkflow.org/

I have developed one workflow for steps in OpenRefine.

2.5 Data file and cleaning summary

Farmers market data, and this a summary of some of the columns and cleaning method adopted

Summary of the file:



63 Columns 8,740 Rows 8 Data Types

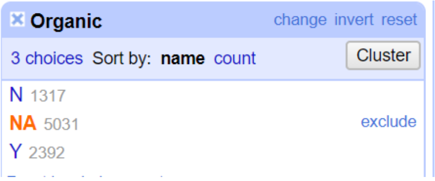
Column Name	Description	Cleaning Operations
FMID	ID farmer Market	Data is clean no operations on this column
MarketName	Name of the Market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization.
Website	Website URL	Open Refine:

		1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization. Example: http://www.seela.org http://www.seela.org/ Standardized value: http://www.seela.org Trifacta: Check mismatch of column value formats
Facebook	Facebook link	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization. Trifacta: Check mismatch of column value formats
Twitter	Twitter Username	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization Not available information is converted to space or null in open refine. n/a(7 rows) N/A(2 rows)

		NA(2 rows) Converted to space
YouTube	Youtube Channel	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization This info is missing in the file so standardized to NA
OtherMedia	Other social media accounts	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and ngram fingerprint data standardization Ngram size 2
Street	Street address for store	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
City	City location for store	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Upper case for city name 3.Cluster and edit using key collision and finger print for data standardization

		Created city clean column
County	Name of county	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
State	State location for farmers market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Zip	Zip code for farmers market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces No clustering is required for this column
Season1Date	First season date for maket	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Season1Time	First season time for market	Trim white spaces and replace new lines in OpenRefine

Season2Date	Second season date for market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Season2Time	Second season time for market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Season3Date	Third season date for market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Season3Time	Third season time for market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Season4Date	Fourth season date for market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization

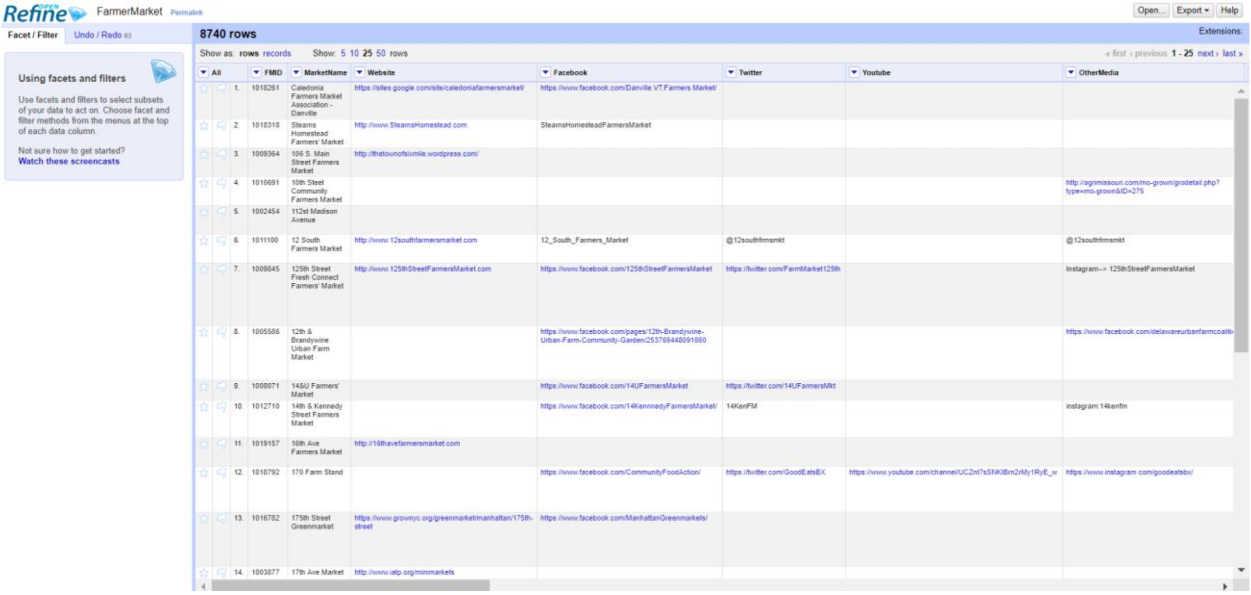
Season4Time	Fourth season time for market	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces 2.Cluster and edit using key collision and finger print for data standardization
Location	Location of the market	Data was clean for this
Organic	Flag to indicate if store sells organic products or not	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces Changed facet – to NA to identify if FLAG is NA i.e. Not Available.  Organic change invert reset 3 choices Sort by: name count Cluster N 1317 NA 5031 exclude Y 2392
updateTime	Time the info was updated	Open Refine: 1.Trim white spaces and collapse consecutive whitespaces

3. Data Cleaning with Open Refine and Trifacta tool set

Open Refine tool was used to do data cleaning. Common transforms were used to remove white spaces and to collapse consecutive white spaces. Clustering was done to do data standardization for fields.

Trifecta data wrangler was used for data wrangling and cleanup individual fields.

Sample screenshot for the openrefine operations:



8740 rows

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

« first ‹ previous 1 - 25 next › last »

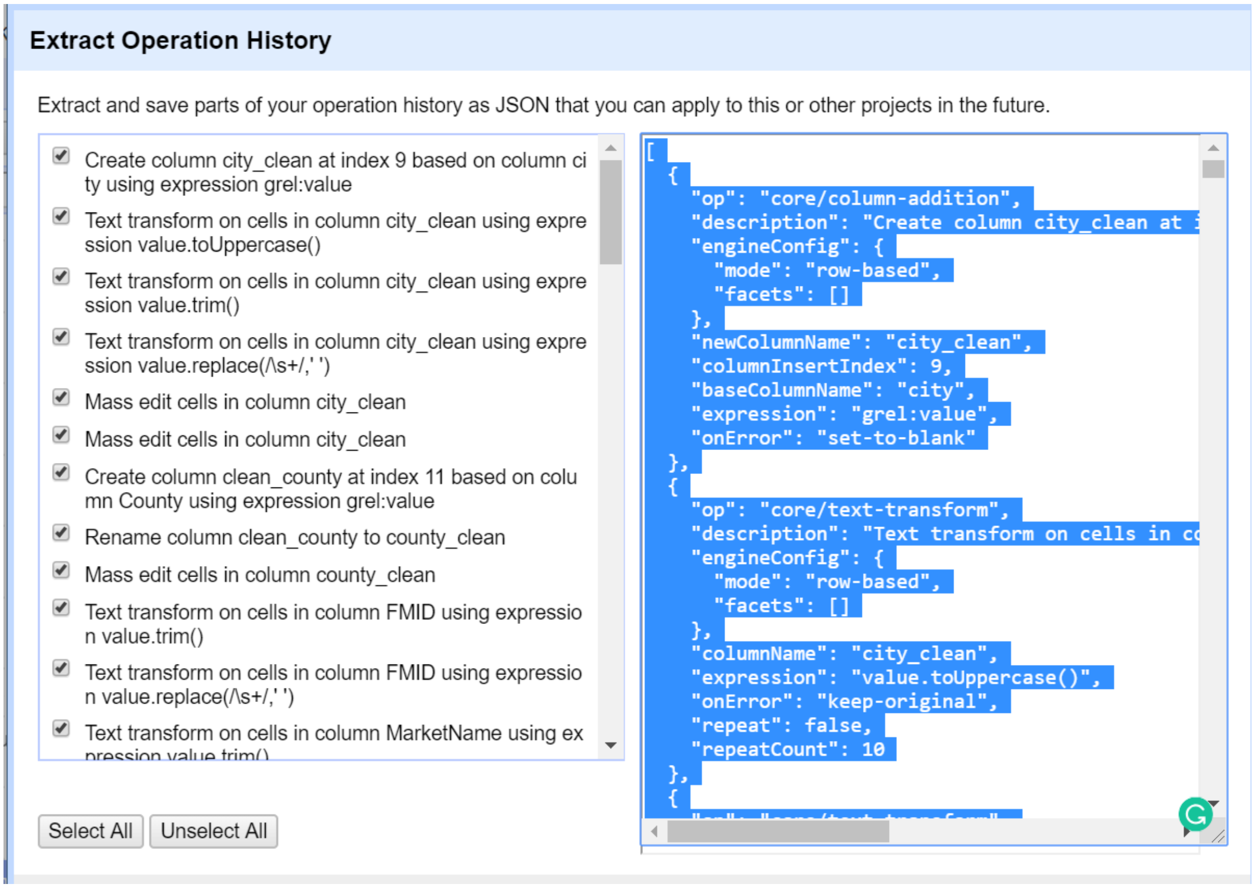
Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

All	FMID	MarketName	Website	Facebook	Twitter	Youtube	OtherMedia
1.	1018261	Caledonia Farmers Market Association - Danville	https://sites.google.com/site/caledoniafarmersmarket/	https://www.facebook.com/Danville.VT.Farmers.Market/			
2.	1018318	Stearns Homestead Farmers' Market	http://www.StearnsHomestead.com	StearnsHomesteadFarmersMarket			
3.	1009364	106 S. Main Street Farmers Market	http://thetownofsixmile.wordpress.com/				
4.	1010691	10th Steet Community Farmers Market					http://agrimissouri.com/mo-grown/grodetail.php?type=mo-grown&ID=275
5.	1002454	112st Madison Avenue					
6.	1011100	12 South Farmers Market	http://www.12southfarmersmarket.com	12_South_Farmers_Market	@12southfrmsmkt		@12southfrmsmkt
7.	1009845	125th Street Fresh Connect Farmers' Market	http://www.125thStreetFarmersMarket.com	https://www.facebook.com/125thStreetFarmersMarket	https://twitter.com/FarmMarket125th		Instagram--> 125thStreetFarmersMarket
8.	1005586	12th & Brandywine Urban Farm Market		https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860			https://www.facebook.com/delawareurbanfarmcoaliti
9.	1008071	14&U Farmers' Market		https://www.facebook.com/14UFarmersMarket	https://twitter.com/14UFarmersMkt		
10.	1012710	14th & Kennedy Street Farmers Market		https://www.facebook.com/14KennnedyFarmersMarket/	14KenFM		instagram:14kenfm
11.	1019157	16th Ave Farmers Market	http://16thavefarmersmarket.com				
12.	1018792	170 Farm Stand		https://www.facebook.com/CommunityFoodAction/	https://twitter.com/GoodEatsBX	https://www.youtube.com/channel/UCZnl7sSNKlBm2rMy1RyE_w	https://www.instagram.com/goodeatsbx/
13.	1016782	175th Street Greenmarket	https://www.grownyc.org/greenmarket/manhattan/175th-street	https://www.facebook.com/ManhattanGreenmarkets/			
14.	1003877	17th Ave Market	http://www.iatp.org/minimarkets				

Here is the sample extract for open refine operations done for data quality:



Extract Operation History

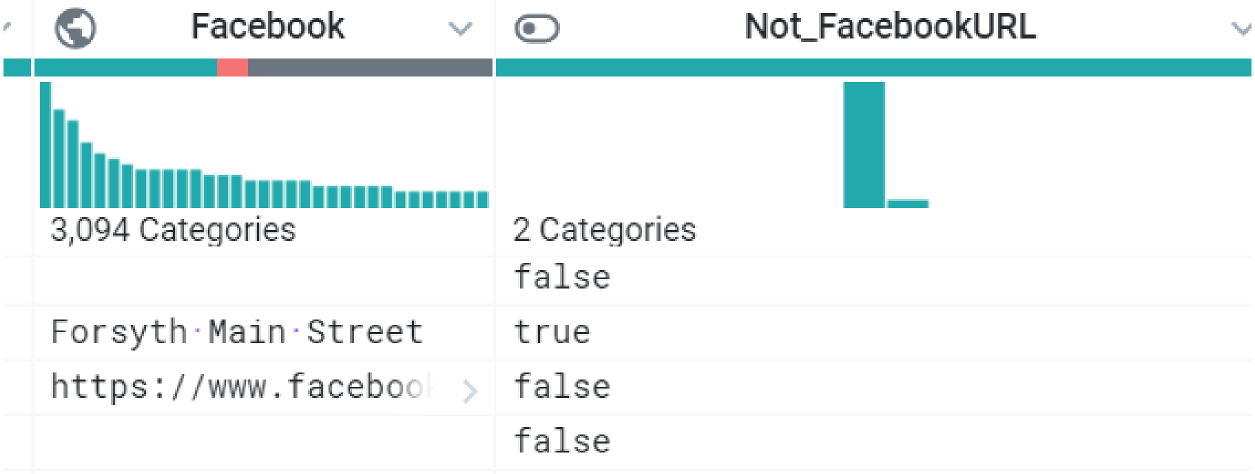
Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

- ☑ Create column city_clean at index 9 based on column city using expression grel:value
- ☑ Text transform on cells in column city_clean using expression value.toUppercase()
- ☑ Text transform on cells in column city_clean using expression value.trim()
- ☑ Text transform on cells in column city_clean using expression value.replace(/\s+/,' ')
- ☑ Mass edit cells in column city_clean
- ☑ Mass edit cells in column city_clean
- ☑ Create column clean_county at index 11 based on column County using expression grel:value
- ☑ Rename column clean_county to county_clean
- ☑ Mass edit cells in column county_clean
- ☑ Text transform on cells in column FMID using expression value.trim()
- ☑ Text transform on cells in column FMID using expression value.replace(/\s+/,' ')
- ☑ Text transform on cells in column MarketName using expression value.trim()

```
[
  {
    "op": "core/column-addition",
    "description": "Create column city_clean at i
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "newColumnName": "city_clean",
    "columnInsertIndex": 9,
    "baseColumnName": "city",
    "expression": "grel:value",
    "onError": "set-to-blank"
  },
  {
    "op": "core/text-transform",
    "description": "Text transform on cells in co
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "city_clean",
    "expression": "value.toUppercase()",
    "onError": "keep-original",
    "repeat": false,
    "repeatCount": 10
  },
  {
```

More details are provided as file in the final deliverables.

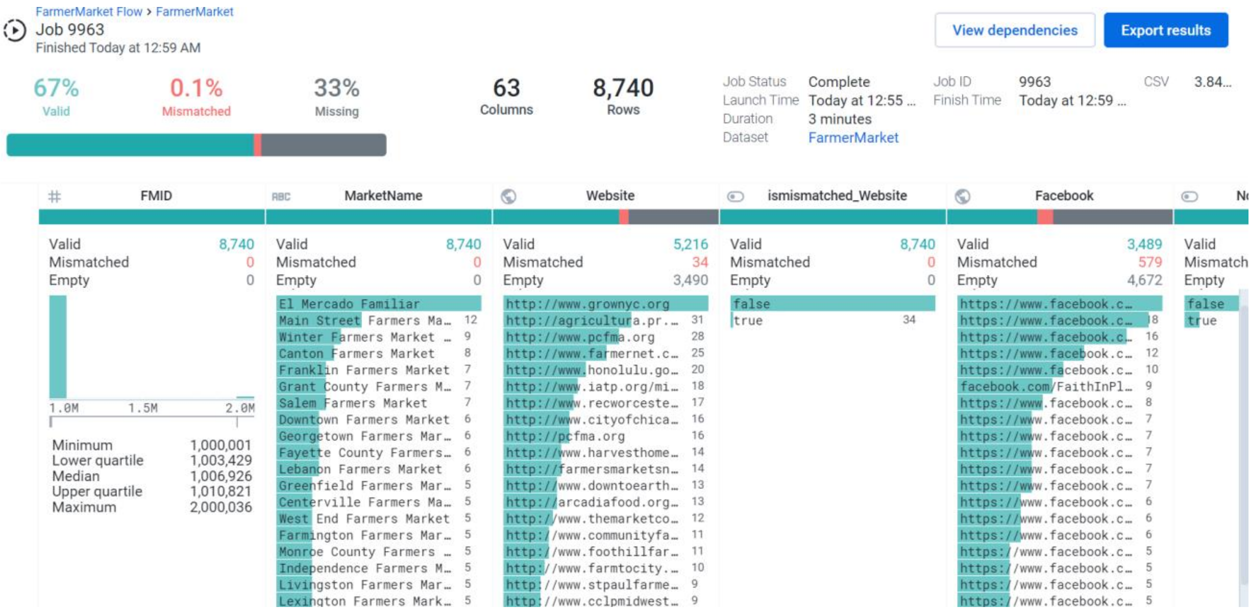
Trifacta data wrangler steps:



Facebook	Not_FacebookURL
3,094 Categories	2 Categories
	false
Forsyth Main Street	true
https://www.faceboo	false
	false

Some Facebook column value don’t have URL. They are represented in different column by flag called Not_facebookURL to identify such kind of columns.

Data Profiling in Trifacta Wrangler tool:



FarmerMarket Flow > FarmerMarket
Job 9963
Finished Today at 12:59 AM

67%	0.1%	33%	63	8,740
Valid	Mismatched	Missing	Columns	Rows

Job Status	Complete	Job ID	9963	CSV	3.84…
Launch Time	Today at 12:55 …	Finish Time	Today at 12:59 …		
Duration	3 minutes				
Dataset	FarmerMarket				

FMID

Valid	8,740
Mismatched	0
Empty	0
Minimum	1,000,001
Lower quartile	1,003,429
Median	1,006,926
Upper quartile	1,010,821
Maximum	2,000,036

MarketName

Valid	8,740
Mismatched	0
Empty	0
El Mercado Familiar	
Main Street Farmers Ma…	12
Winter Farmers Market …	9
Canton Farmers Market	8
Franklin Farmers Market	7
Grant County Farmers M…	7
Salem Farmers Market	7
Downtown Farmers Market	6
Georgetown Farmers Mar…	6
Fayette County Farmers…	6
Lebanon Farmers Market	6
Greenfield Farmers Mar…	5
Centerville Farmers Ma…	5
West End Farmers Market	5
Farmington Farmers Mar…	5
Monroe County Farmers …	5
Independence Farmers M…	5
Livingston Farmers Mar…	5
Lexington Farmers Mark…	5

Website

Valid	5,216
Mismatched	34
Empty	3,490
http://www.grownyc.org	
http://agricultura.pr.…	31
http://www.pcfma.org	28
http://www.farmernet.c…	25
http://www.honolulu.go…	20
http://www.iatp.org/mi…	18
http://www.recworceste…	17
http://www.cityofchica…	16
http://pcfma.org	16
http://www.harvesthome…	14
http://farmersmarketsn…	14
http://www.downtoearth…	13
http://arcadiafood.org…	13
http://www.themarketco…	12
http://www.communityfa…	11
http://www.foothillfar…	11
http://www.farmtocity.…	10
http://www.stpaulfarme…	9
http://www.cclpmidwest…	9

ismismatched_Website

Valid	8,740
Mismatched	0
Empty	0
false	
true	34

Facebook

Valid	3,489
Mismatched	579
Empty	4,672
https://www.facebook.c…	
https://www.facebook.c…	18
https://www.facebook.c…	16
https://www.facebook.c…	12
https://www.facebook.c…	10
facebook.com/FaithInPl…	9
https://www.facebook.c…	8
https://www.facebook.c…	7
https://www.facebook.c…	7
https://www.facebook.c…	7
https://www.facebook.c…	7
https://www.facebook.c…	7
https://www.facebook.c…	6
https://www.facebook.c…	6
https://www.facebook.c…	6
https://www.facebook.c…	5
https://www.facebook.c…	5
https://www.facebook.c…	5
https://www.facebook.c…	5

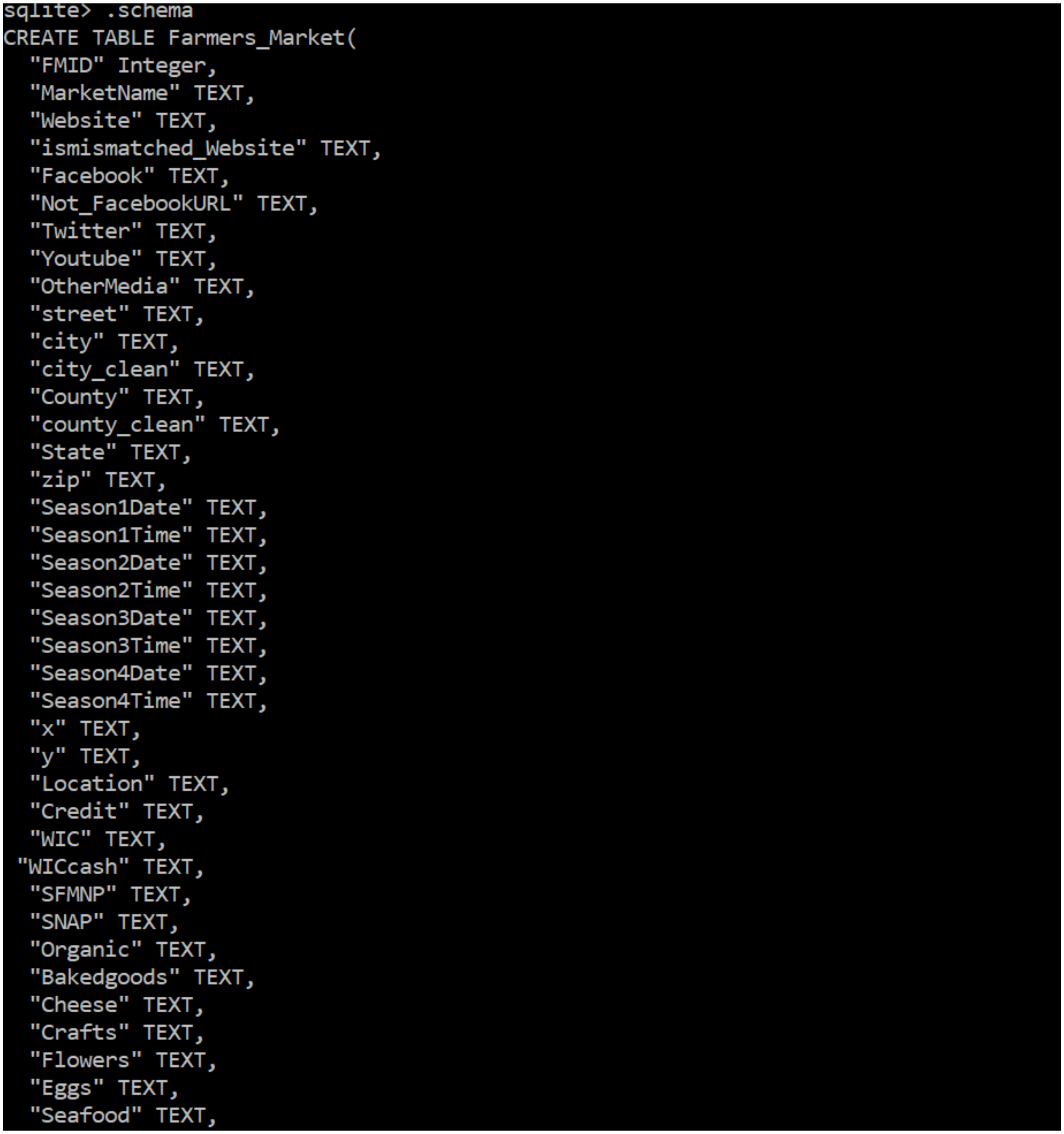
N

Valid
Mismatch
Empty
false
true

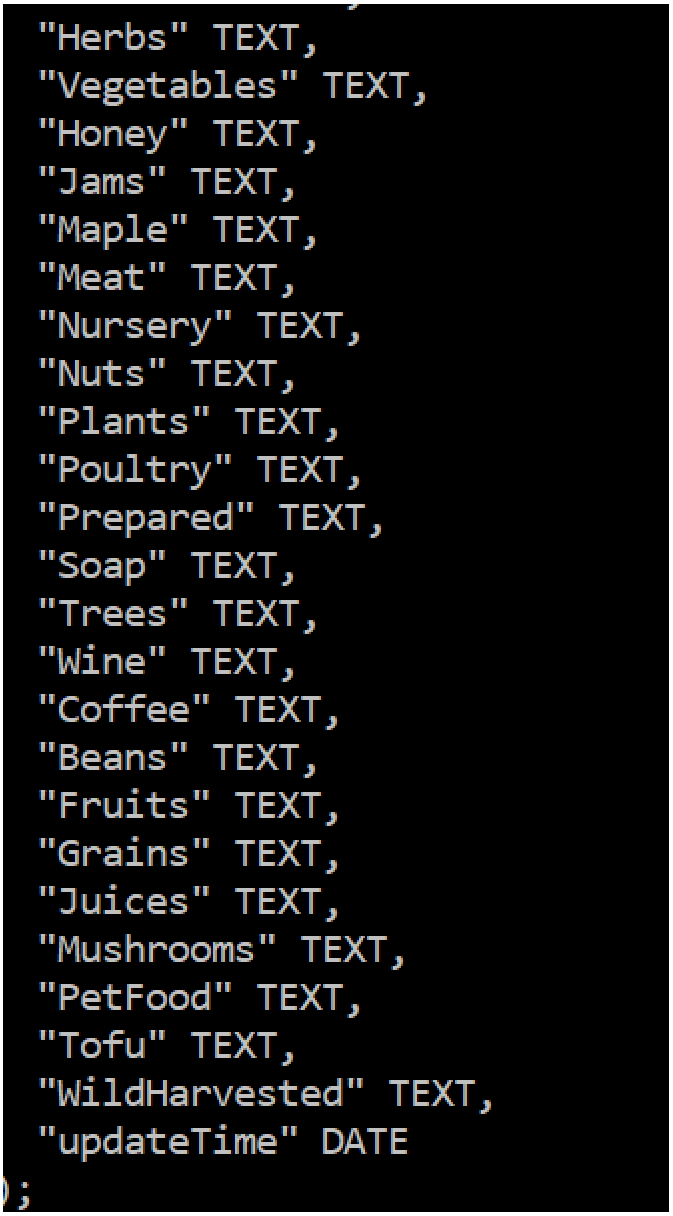
4. Relation Database Schema

I. Schema

Sqlite3 was used to load the data file exported from the trifacta data wrangler for next step in the project. Screenshot of the schema:



```
sqlite> .schema
CREATE TABLE Farmers_Market(
  "FMID" Integer,
  "MarketName" TEXT,
  "Website" TEXT,
  "ismismatched_Website" TEXT,
  "Facebook" TEXT,
  "Not_FacebookURL" TEXT,
  "Twitter" TEXT,
  "Youtube" TEXT,
  "OtherMedia" TEXT,
  "street" TEXT,
  "city" TEXT,
  "city_clean" TEXT,
  "County" TEXT,
  "county_clean" TEXT,
  "State" TEXT,
  "zip" TEXT,
  "Season1Date" TEXT,
  "Season1Time" TEXT,
  "Season2Date" TEXT,
  "Season2Time" TEXT,
  "Season3Date" TEXT,
  "Season3Time" TEXT,
  "Season4Date" TEXT,
  "Season4Time" TEXT,
  "x" TEXT,
  "y" TEXT,
  "Location" TEXT,
  "Credit" TEXT,
  "WIC" TEXT,
 "WICcash" TEXT,
  "SFMNP" TEXT,
  "SNAP" TEXT,
  "Organic" TEXT,
  "Bakedgoods" TEXT,
  "Cheese" TEXT,
  "Crafts" TEXT,
  "Flowers" TEXT,
  "Eggs" TEXT,
  "Seafood" TEXT,
```



```
  "Herbs" TEXT,
  "Vegetables" TEXT,
  "Honey" TEXT,
  "Jams" TEXT,
  "Maple" TEXT,
  "Meat" TEXT,
  "Nursery" TEXT,
  "Nuts" TEXT,
  "Plants" TEXT,
  "Poultry" TEXT,
  "Prepared" TEXT,
  "Soap" TEXT,
  "Trees" TEXT,
  "Wine" TEXT,
  "Coffee" TEXT,
  "Beans" TEXT,
  "Fruits" TEXT,
  "Grains" TEXT,
  "Juices" TEXT,
  "Mushrooms" TEXT,
  "PetFood" TEXT,
  "Tofu" TEXT,
  "WildHarvested" TEXT,
  "updateTime" DATE
);
```

II. Integrity Constraints

Following integrity constraints where identified and described here:

Data Integrity Constraints:

Basic data sanity check:

- ✓ Check number of rows loaded into the sqlite3. Please note If the table already exists, the sqlite3 tool uses all the rows, including the first row, in the CSV file as the actual data to import. Therefore while matching data rows I have deleted the first row from the file as its metadata and not the actual data.

Following IC are written in denial form:

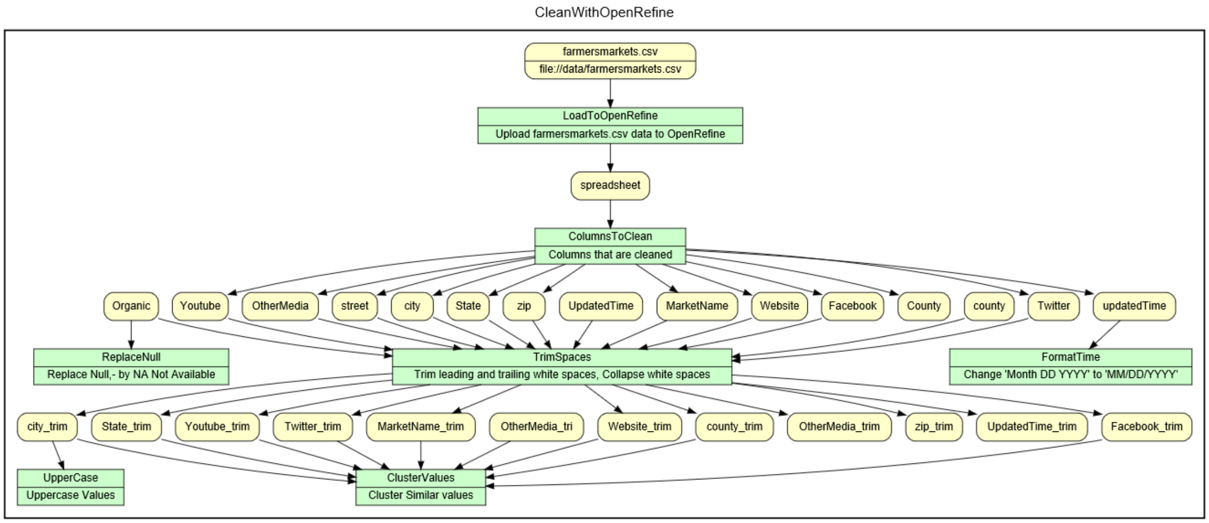
- ✓ Every market has a website URL.
- ✓ URL format is consistent for farmer's market
- ✓ Every record has city and county information.

5. Workflow Model

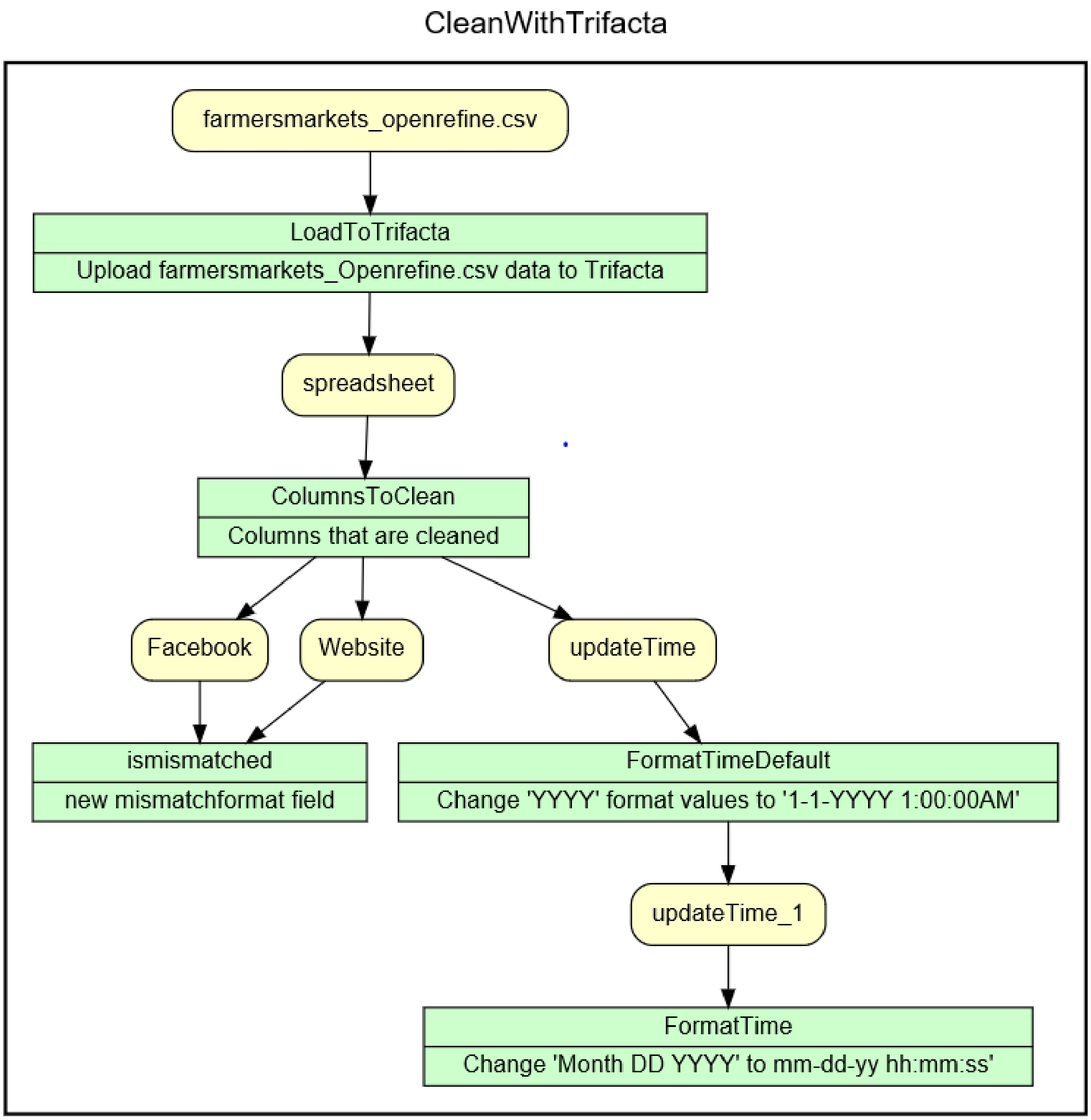
I have used YesWorkflow to create a provenance representation of the data cleaning process of the farmers market data. I have used http://try.yesworkflow.org/ to create the workflow with the web-based version which allows us to see our changes as they are made.

I have developed one workflow focusing on the two main steps used to clean the data, OpenRefine and Trifacta. Within those two main steps, there are numerous sub-steps to show the exact steps taken to clean the steps.

Here is the flow diagram for open refine:



Trifacta:



Recipe example:

'''

@begin CleanWithOpenRefine @desc OpenRefine Workflow for farmersmarkets dataset

@in farmersmarkets.csv @uri file://data/farmersmarkets.csv

@begin LoadToOpenRefine @desc Upload farmersmarkets.csv data to OpenRefine

@in farmersmarkets.csv @uri file://data/farmersmarkets.cv

@out spreadsheet

@end CleanWithOpenRefine

@begin ColumnsToClean @desc Columns that are cleaned

@in spreadsheet

@out MarketName

@out Website

@out Facebook

@out Twitter

@out Youtube

@out OtherMedia

@out street

@out city

@out County

@out State

@out zip

@out updatedTime

@end ColumnsToClean

@begin TrimSpaces @desc Trim leading and trailing white spaces, Collapse white spaces

@in MarketName

@in Website

@in Facebook

```
@in Twitter

@in Youtube

@in OtherMedia

@in street

@in city

@in county

@in State

@in zip

@in UpdatedTime

@out MarketName_trim

@out Website_trim

@out Facebook_trim

@out Twitter_trim

@out Youtube_trim

@out OtherMedia_trim

@out city_trim

@out county_trim

@out State_trim

@out zip_trim

@in Organic

@out UpdatedTime_trim

@end TrimSpaces
```

@begin FormatTime @desc Change 'Month DD YYYY' to 'MM/DD/YYYY'

@in updatedTime

@end FormatTime

@begin ClusterValues @desc Cluster Similar values

@in MarketName_trim

@in city_trim

@in Website_trim

@in county_trim

@in Facebook_trim

@in city_trim

@in State_trim

@in Youtube_trim

@in Twitter_trim

@in OtherMedia_trim

@end ClusterValues

@begin UpperCase @desc Uppercase Values

@in city_trim

@end DUpperCase

@begin ReplaceNull @desc Replace Null,- by NA Not Available

@in Organic

@end ReplaceNull

```
@out farmersmarkets-or.csv @uri file://data/farmersmarkets-or.csv
@end CleanWithOpenRefine
'''
'''
@begin CleanWithTrifacta @desc Trifacta Workflow for farmersmarkets dataset
@in farmersmarkets.csv @uri file://data/farmersmarkets.csv

  @begin LoadToTrifacta @desc Upload farmersmarkets_Openrefine.csv data to Trifacta
  @in farmersmarkets_openrefine.csv @uri file://data/farmersmarketsopenrefine.cv
  @out spreadsheet
  @end CleanWithTrifacta

  @begin ColumnsToClean @desc Columns that are cleaned
  @in spreadsheet
  @out Facebook
  @out Website
  @out updateTime
  @end ColumnsToClean

  @begin ismismatched @desc new mismatchformat field
  @in Facebook
  @in Website
  @end ismismatched
```

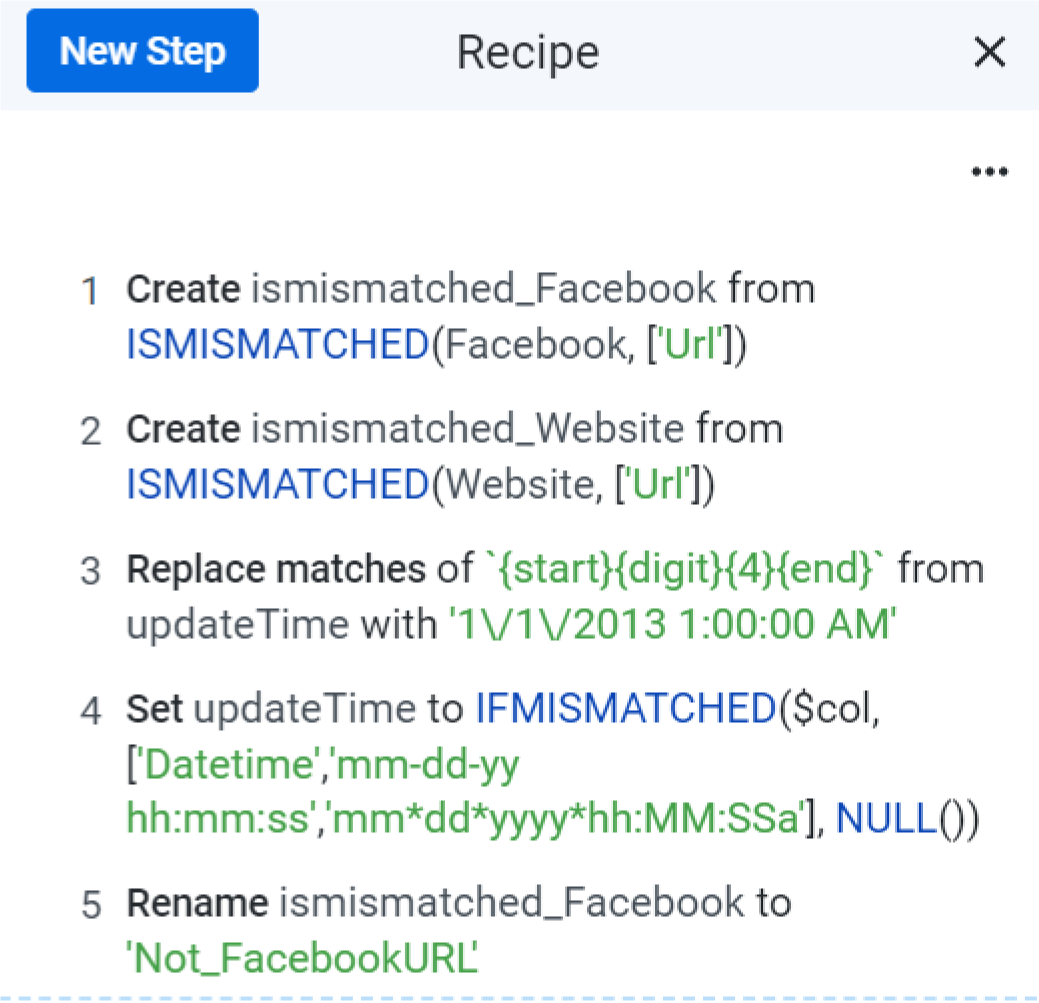
```
@begin FormatTimeDefault @desc Change 'YYYY' format values to '1-1-YYYY 1:00:00AM'
  @in updateTime
  @out updateTime_1
  @end FormatTimeDefault

  @begin FormatTime @desc Change 'Month DD YYYY' to mm-dd-yy hh:mm:ss'
  @in updateTime_1

  @end FormatTime

@out farmersmarkets-or.csv @uri file://data/farmersmarkets-or.csv
@end CleanWithOpenRefine
'''
```

Appendix A – Trifacta Recipe



1 **Create** ismismatched_Facebook from ISMISMATCHED(Facebook, ['Url'])

2 **Create** ismismatched_Website from ISMISMATCHED(Website, ['Url'])

3 **Replace matches** of `{start}{digit}{4}{end}` from updateTime with '1\/1\/2013 1:00:00 AM'

4 **Set** updateTime to IFMISMATCHED($col, ['Datetime','mm-dd-yy hh:mm:ss','mm*dd*yyyy*hh:MM:SSa'], NULL())

5 **Rename** ismismatched_Facebook to 'Not_FacebookURL'

Appendix B – SQLite Logical Integrity Checking Script and Output

1.Every market has a website URL

Not all farmer's market have website

```
select count(*) from farmers_market where Website="" ;
```

```
3490
```

2.IC check website validity

In negate form finds the non compliant number

```
select count(*) from farmers_market where not_FacebookURL="TRUE";
564
```

3.Every record has city and county information

```
sqlite> select count(*) from farmers_market where city="" OR county="";
523
```

Appendix C – YesWorkflow Script

This is the sample workflow text, more details on the github page[2]

```
'''
@begin CleanWithOpenRefine @desc OpenRefine Workflow for farmersmarkets dataset
@in farmersmarkets.csv @uri file://data/farmersmarkets.csv

  @begin LoadToOpenRefine @desc Upload farmersmarkets.csv data to OpenRefine
  @in farmersmarkets.csv @uri file://data/farmersmarkets.cv
  @out spreadsheet
  @end CleanWithOpenRefine

  @begin ColumnsToClean @desc Columns that are cleaned
  @in spreadsheet
  @out MarketName
  @out Website
  @out Facebook
  @out Twitter
  @out Youtube
  @out OtherMedia
  @out street
  @out city
  @out County
  @out State
  @out zip
  @out updatedTime
  @end ColumnsToClean

  @begin TrimSpaces @desc Trim leading and trailing white space
  @in MarketName
```

```
@in Website
@in Facebook
@in Twitter
@in Youtube
@in OtherMedia
@in street
@in city
@in county
@in State
@in zip
@in UpdatedTime
@out MarketName_trim
@out Website_trim
@out Youtube_trim
@out Season1Date_trim
@end TrimSpaces


@begin CollapseSpaces @desc Collapse consecutive white spaces
@in MarketName
@in Facebook
@in OtherMeida
@in street
@in city
@in Season1Date
@out MarketName_o
@out city_
@out Facebook_
```

@out Season1Date_

@end CollapseSpaces

@begin FormatTime @desc Change 'Month DD YYYY' to 'MM/DD/YYYY'

@in updatedTime

@end FormatTime

@begin ClusterValues @desc Cluster Similar values

@in MarketName_

@in city_

@end ClusterValues

@begin DeleteNAorNONE @desc Delete na or none values

@in Facebook_

@in Twitter_

@in Youtube_

@end DeleteNAorNONE

@begin SplitInto2Columns @desc Split date into Start Date and End Date

@in Season1Date_

@out Season1StartDate

@out Season1EndDate

@end SPlitInto2Columns

@begin CleanFormatZip @desc Clean any string and reformat to NNNNN or NNNNN-NNNN

@in zip

@end CleanFormatZip

```
 @begin ToDateType @desc Change data type to date

 @in Season1StartDate

 @in Season1EndDate

 @end ToDateType


@out farmersmarkets-or.csv @uri file://data/farmersmarkets-or.csv

@end CleanWithOpenRefine

'''
```

Appendix D – References and Github URL link

1. https://www.ams.usda.gov/local-food-directories/farmersmarkets
2. https://github.com/varunuiuc/DataCleaning