**Project Report –CS513 Summer 2020**

**Dataset: US Farmers Markets**

**Saurav Chetry (**[**schetry2@illiois.edu**](mailto:schetry2@illiois.edu)**)**

1. **Overview & Initial Assessment of Dataset:**

Data Description:

The data is about farmers market, the produce available at different seasons and the location co-ordinates to the markets for customers to plan their visit for purchases. Besides, there is also information about what types of cash transactions are offered by individual markets. In short it looks like a Google Maps review page of a map where such information is available for end users.

There are total Rows: 8687, Columns: 59 and contains the market name, website, address, social media information, products, season times, and updated time for the entry.

Feasible Use-Case and Data Cleaning Goals:

The data can be used for an online search tool for customers looking to purchase local produce. With the name, location, season data and time, produce details, purchasing options and location co-ordinates, customers can use the tool to navigate to the market.

For the navigation use case, the markets must offer at least either the latitude longitude co-ordinates or a combination of any of street, state, county or zip. Besides information about social media can be used to collate the reviews and photos of the markets in the search use case.

Social Media and Location information must be cleaned and made ready to use for Navigation and Search use-case.

This will involve correcting Social Media URLs, hyperlinks and market websites.

The ZIP code needs to be correct by removing special characters which are not required.

Season Date/Time must be correct to match the produce availability and to appear on search based on relevance.

Unique Market name and City name will in the use case.

Correct “UpdateTime” is useful for customers to see the latest information.

Purchasing options and produce options are useful for convenience to the customers.

Data Ready for following Use-Cases:

1. Offline database about markets, location and season details. This is a look-up use case good for record-keeping or auditing.

Partially ready Data for following Use-Cases

1. Navigation apps which can locate markets based on latitudes and longitudes.
2. Online application for search, review of markets – data cleaning is required.

Data Not Suitable for following Use-Cases:

1. Contact and Appointment Scheduling apps. No customer contact details available.
2. Legal, administrative solutions as there is no Owner details available.

List of data quality issues:

|  |  |
| --- | --- |
| **Fields** | **Issues** |
| **FMID** | They are unique values. They have whitespaces and trailing/leading zeros. |
| **Market Names** | They have different name variations for the same markets. Same markets do open at different seasons or  years. The names can be clustered to a more suitable name. |
| **SocialMedia (All)** | There issues are incorrect URLs, hyperlinks etc.  For “website”, some fields are blank, and some have both https and http format.  For “Twitter”, some have https, and some have @ format.  For “Facebook”, some do not have the URLs but simple  For “Youtube”, most fields are empty.  For “OtherMedia”, most fields are blank. Incorrect URLs can be removed. |
| **Season1 Date** | There are different formats of dates. MMM, MM-DD-YYYY, YYYY-MM-DD etc.  Some Season Start Dates are Older than the Season End Dates.  Some markets do not have year values for seasonal data, and some are missing end data values.  Some markets do not have year values for seasonal data, and some are missing end data values.  Most products offered in markets have availability for single season. For them other Season Date and Time are missing. |
| **Other Season Dates** | Same issues as Season1 Date |
| **Season Time** | They have whitespaces and trailing/leading zeros. |
| **Street** | They have whitespaces and trailing/leading zeros. |
| **State** | They have whitespaces and trailing/leading zeros. |
| **County** | They have whitespaces and trailing/leading zeros.  There are multiple markets for same city, this can be clustered. |
| **ZIP** | They have non-numeric entries.  Some values are missing. Not a big problem if other details like street and county data are available. |
| **City** | They have whitespaces and trailing/leading zeros. |
| **Location** | Most markets do not have this information. Others have whitespaces and trailing/leading zeros. |
| **Produce (All)** | “Organic” products field has special characters |
| **Update Time** | They have whitespaces and trailing/leading zeros.  There is a mix of “MM/DD/YYYY Time” and “Month DD Year Time” format. |

Overall the data is not useful as-is and would lead to questions and confusions for the end users. The data needs to be cleaned to make it consumption ready.

1. **Data Cleaning with OpenRefine:**

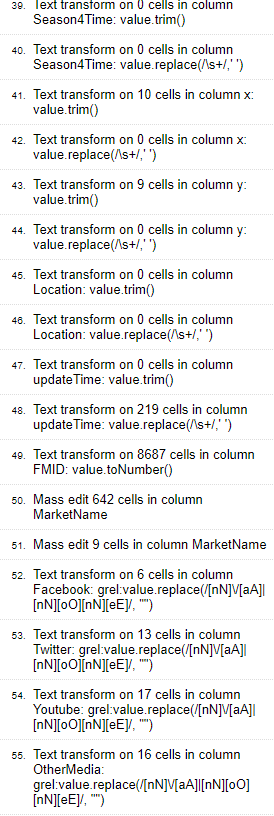
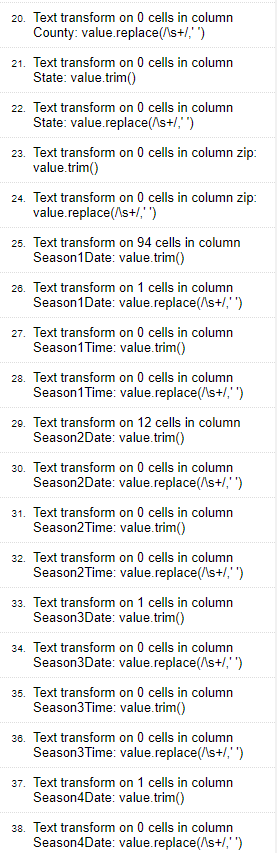
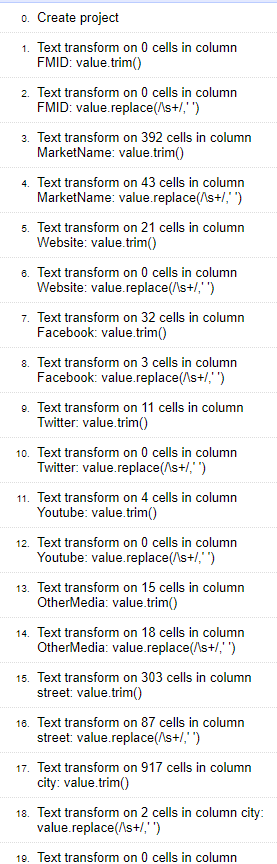
The goal is to implement the use case of the search and navigation application. With that in mind, the following is the overview of the OpenRefine steps carried out on the raw data.

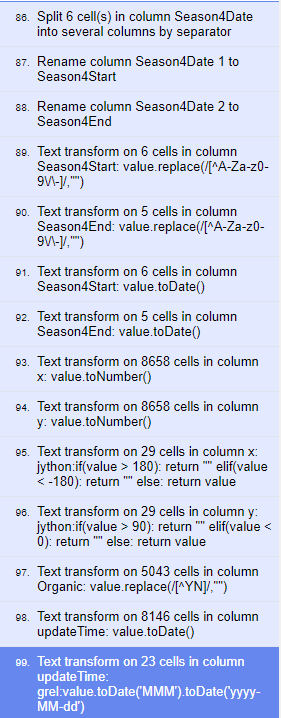
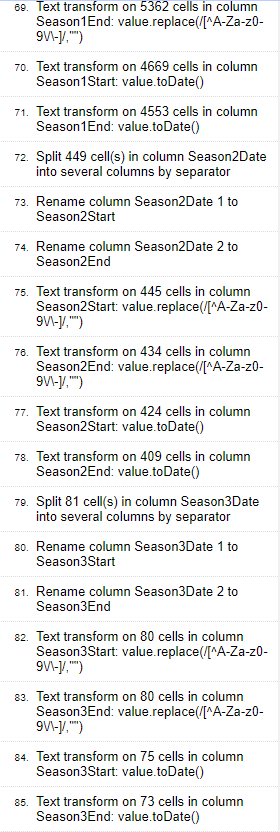
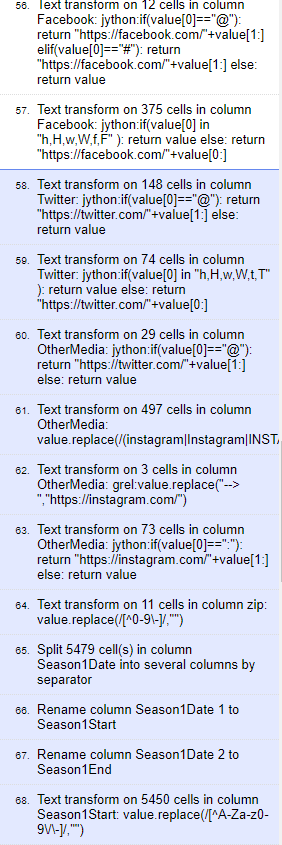
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Operations**  **Column** | **Trim** | **Whitespace** | **Duplicates** | **ToDate** | **Split**  **Columns** | **Value**  **Replace** | **Cluster** | **DataValidation**  **Python/Jython** |
| **FMID** | Checkmark | Checkmark | Checkmark | Close | Close | Close | Close | Close |
| **Market Names** | Checkmark | Checkmark | Checkmark | Close | Close | Close | Checkmark | Close |
| **SocialMedia All** | Checkmark | Checkmark | Checkmark | Close | Close | Checkmark | Close | Checkmark |
| **Season Date** | Checkmark | Checkmark | Checkmark | Checkmark | Checkmark | Close | Close | Close |
| **Season Time** | Checkmark | Checkmark | Close | Close | Close | Close | Close | Close |
| **Street** | Checkmark | Checkmark | Checkmark | Close | Close | Close | Close | Close |
| **State** | Checkmark | Checkmark | Close | Close | Close | Close | Close | Close |
| **County** | Checkmark | Checkmark | Close | Close | Close | Close | Close | Close |
| **ZIP** | Checkmark | Checkmark | Close | Close | Close | Checkmark | Close | Checkmark |
| **City** | Checkmark | Checkmark | Close | Close | Close | Close | Checkmark | Close |
| **Organic** | Checkmark | Checkmark | Close | Close | Close | Checkmark | Close | Checkmark |
| **Update Time** | Checkmark | Checkmark | Close | Checkmark | Close | Close | Close | Close |

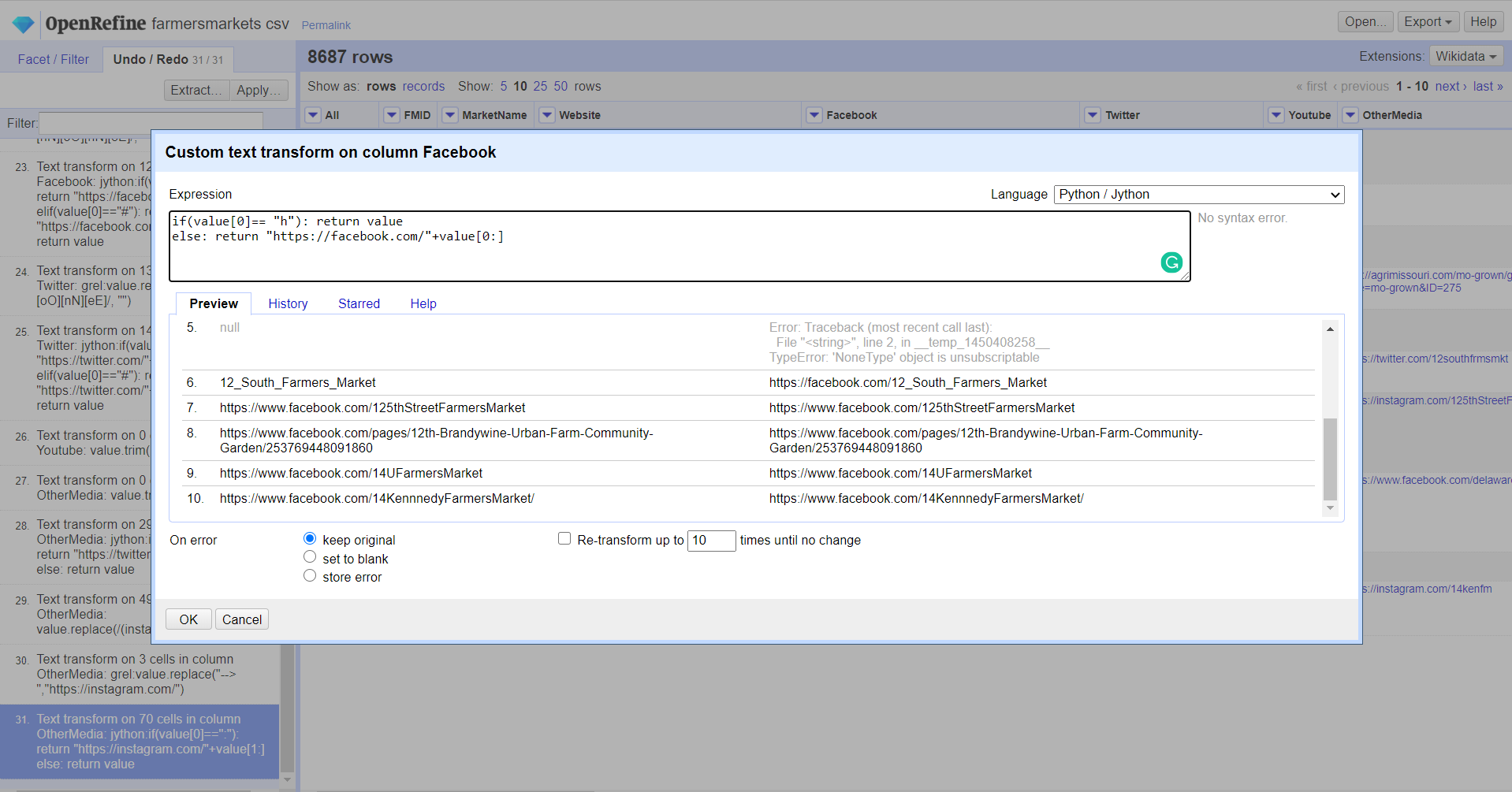
|  |  |
| --- | --- |
| **Fields** | **OpenRefine Operation** |
| **All Fields** | All fields had leading, trailing and intermediate whitespaces. Trimming and collapsing the whitespaces was done on all fields. |
| **FMID,Market Names, Social Media fields, Season Date and Street** | Check for Duplicates. |
| **Market Name** | Market Name had multiple variant names with separate unique FMIDs and other details. The names were clustered, and a common name was chosen for the market.  Make a facet and perform the cluster operation using the \*key-collison\* method and  \*fingerprint\* function. Merge the relevant clusters. |
| **City** | same operation as Market Name |
| **All SeasonDate** | Column split into two with separator “to” to create start and end Season dates.  Text changed to Date format yyyy-MM-dd HH:MM:SS  Remove non-numeric and not allowed characters using **value.replace(/[^A-Za-z0-9\\/\\-]/,\"\")** |
| **All Social Media** | None or NA values were removed using **value.replace(/[nN]\\/[aA]|[nN][oO][nN][eE]/, \"\")**  Texts starting with “@” were replaced with appropriate URLs using Pyhton E.g:  **if(value[0]==\"@\"): return \"https://facebook.com/\"+value[1:]\nelif(value[0]==\"#\"): return \"https://facebook.com/\"+value[1:]\nelse: return value**  Texts not starting with www, WWW, http, Http etc. were replaced with appropriate URLs using Python, e.g.  **if(value[0] in \"h,H,w,W,f,F\" ): return value \nelse: return \"https://facebook.com/\"+value[0:]"**  **if(value[0] in \"h,H,w,W,t,T\" ): return value \nelse: return \"https://twitter.com/\"+value[0:]** |
| **Other Media** | None or NA values were removed using **value.replace(/[nN]\\/[aA]|[nN][oO][nN][eE]/, \"\")**  Text starting with @ was replaced with twitter URL:  **if(value[0]==\"@\"): return \"https://twitter.com/\"+value[1:]\nelse: return value**  Text having mention of Instagram, but no clickable links were replaced as:  **value.replace(/(instagram|Instagram|INSTAGRAM)/,\"\")**  **value.replace(\"--> \",\"https://instagram.com/\")**  **if(value[0]==\":\"): return \"https://instagram.com/\"+value[1:]\nelse: return value** |
| **ZIP** | Invalid and not allowed characters were removed as **value.replace(/[^0-9\\-]/,\"\")** |
| **X, Y** | Converted all text to Numbers  Validated and removed values of **180<X<-180** as:  **if(value > 180): return \"\"\nelif(value < -180): return \"\"\nelse: return value**  Validated and removed values of **0<Y<90** as:  **if(value > 90): return \"\"\nelif(value < 0): return \"\"\nelse: return value** |
| **Organic** | Replaced characters other than Y or N as:  **value.replace(/[^YN]/,\"\")** |
| **UpdateTime** | Text changed to Date format yyyy-MM-dd HH:MM:SS |

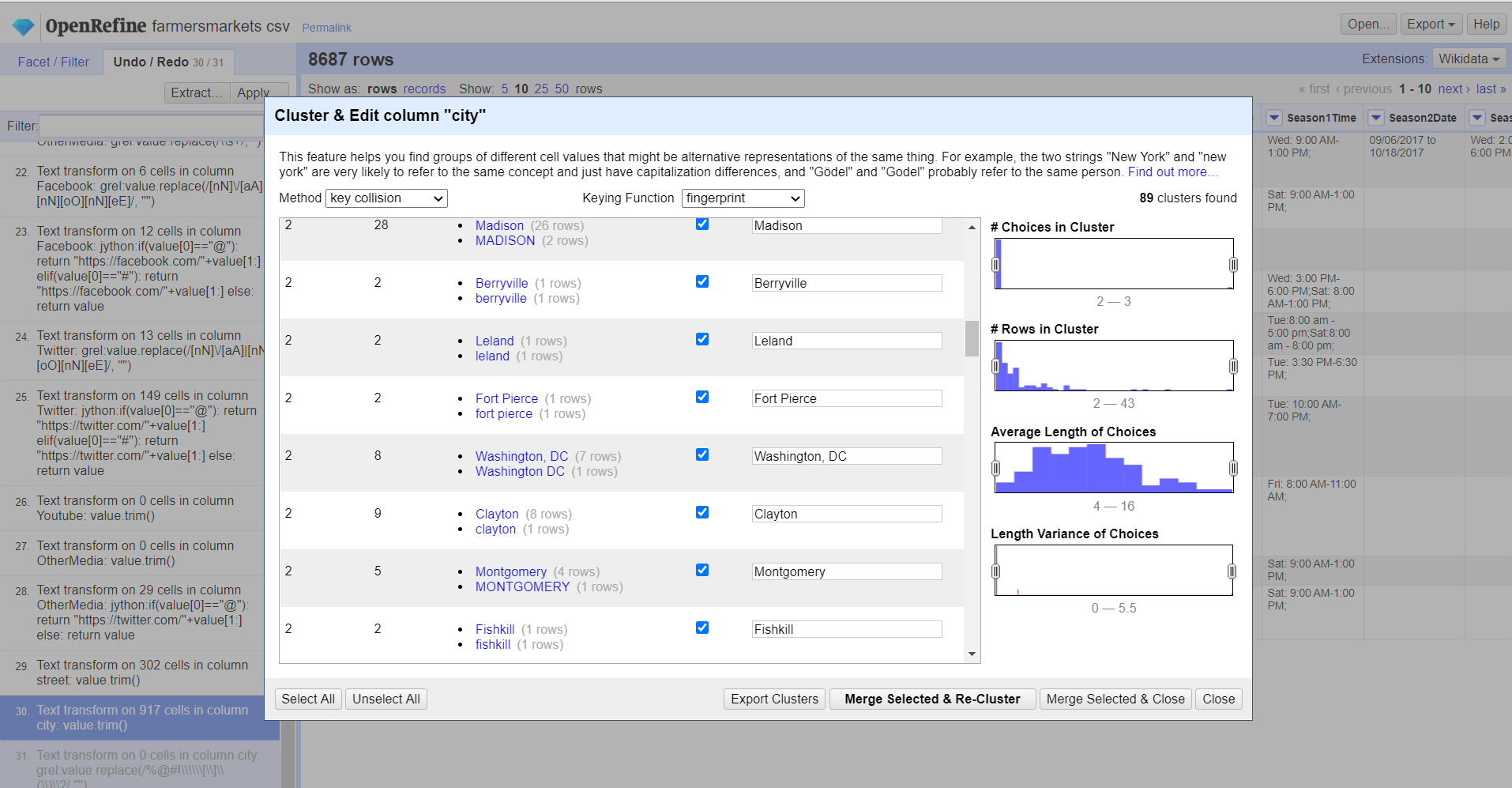
Quantification of above changes:

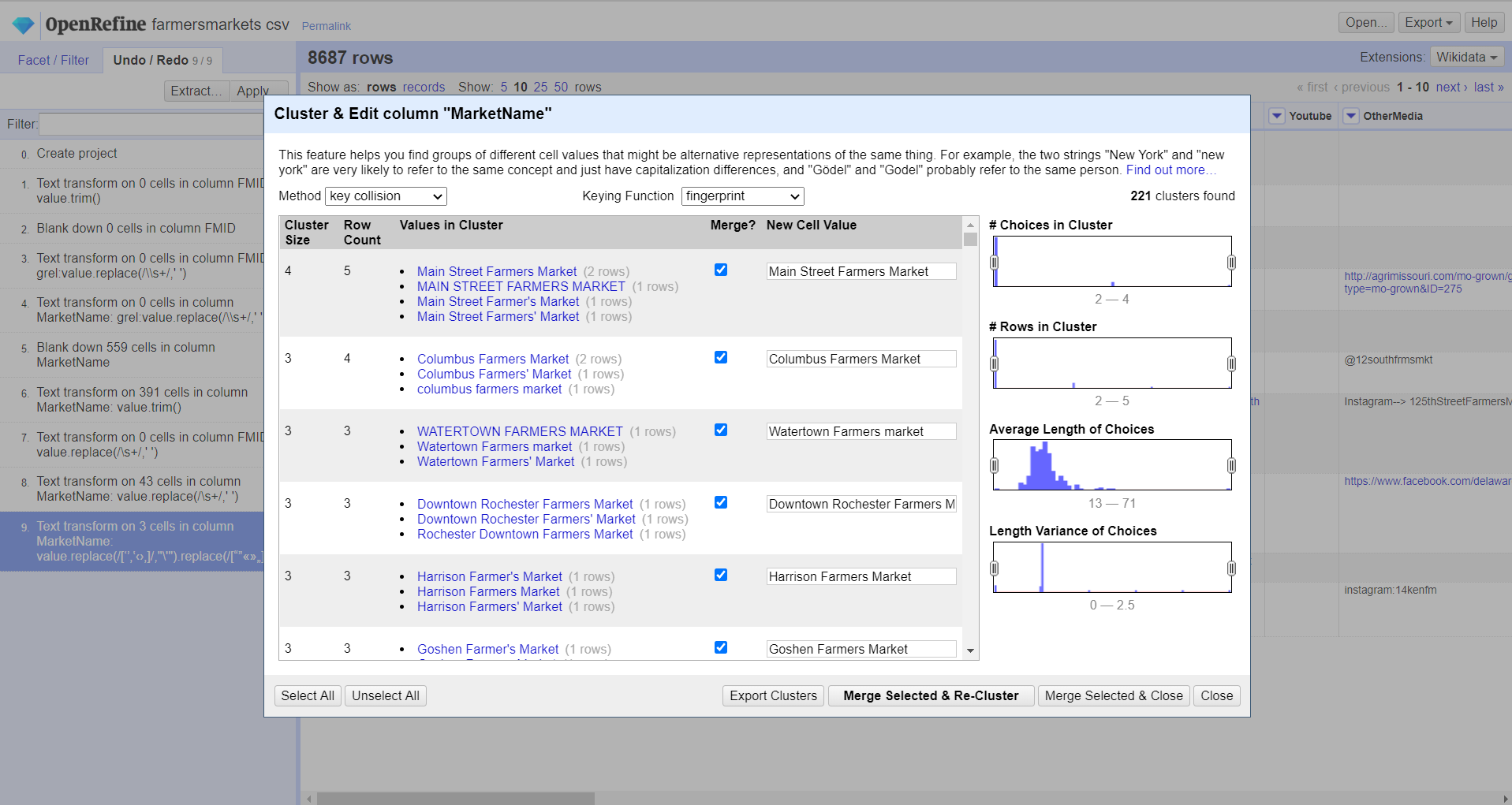
Below are the snapshots of changes effected to the raw data by each OpenRefine operation mentioned above.

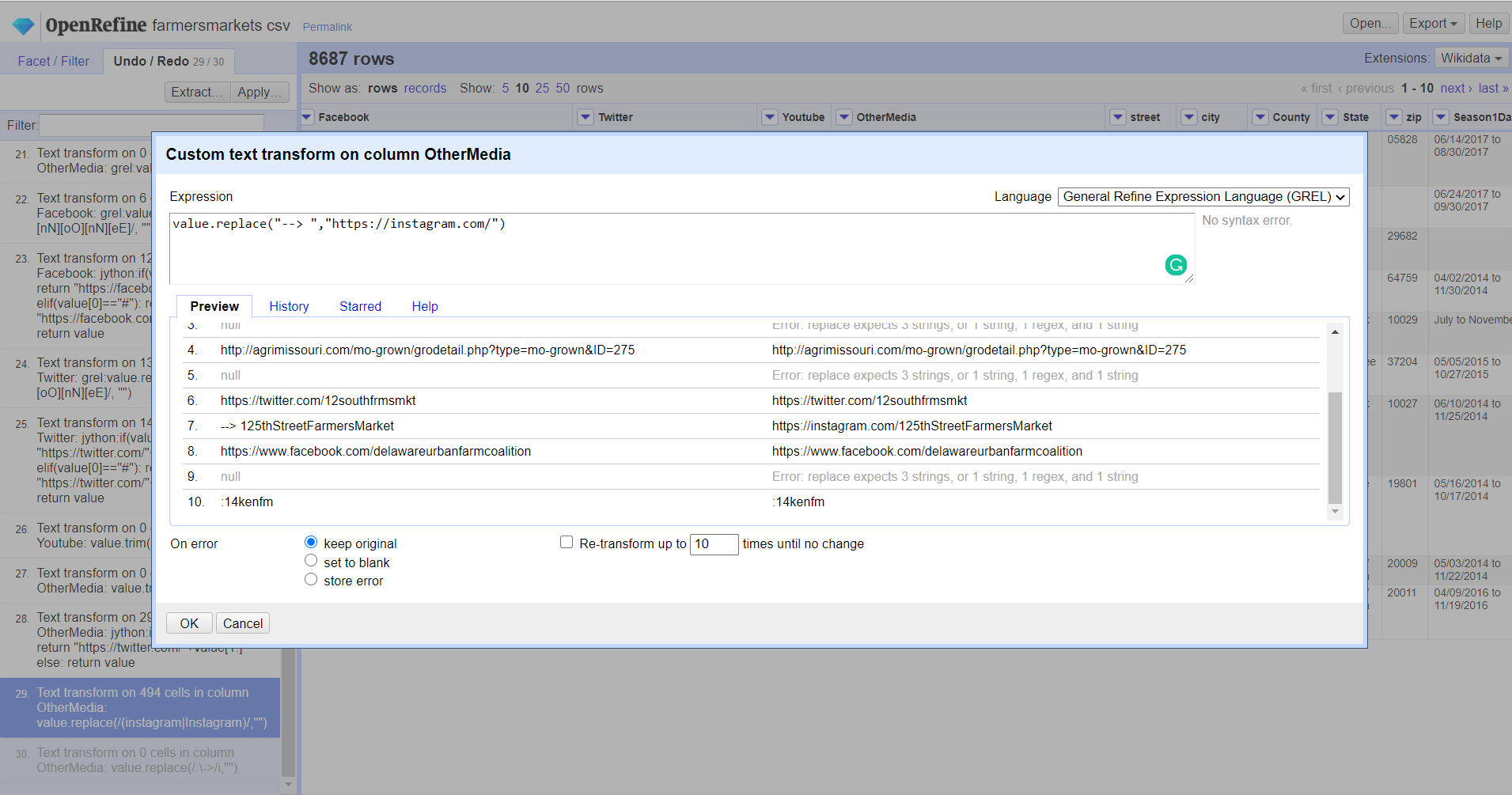


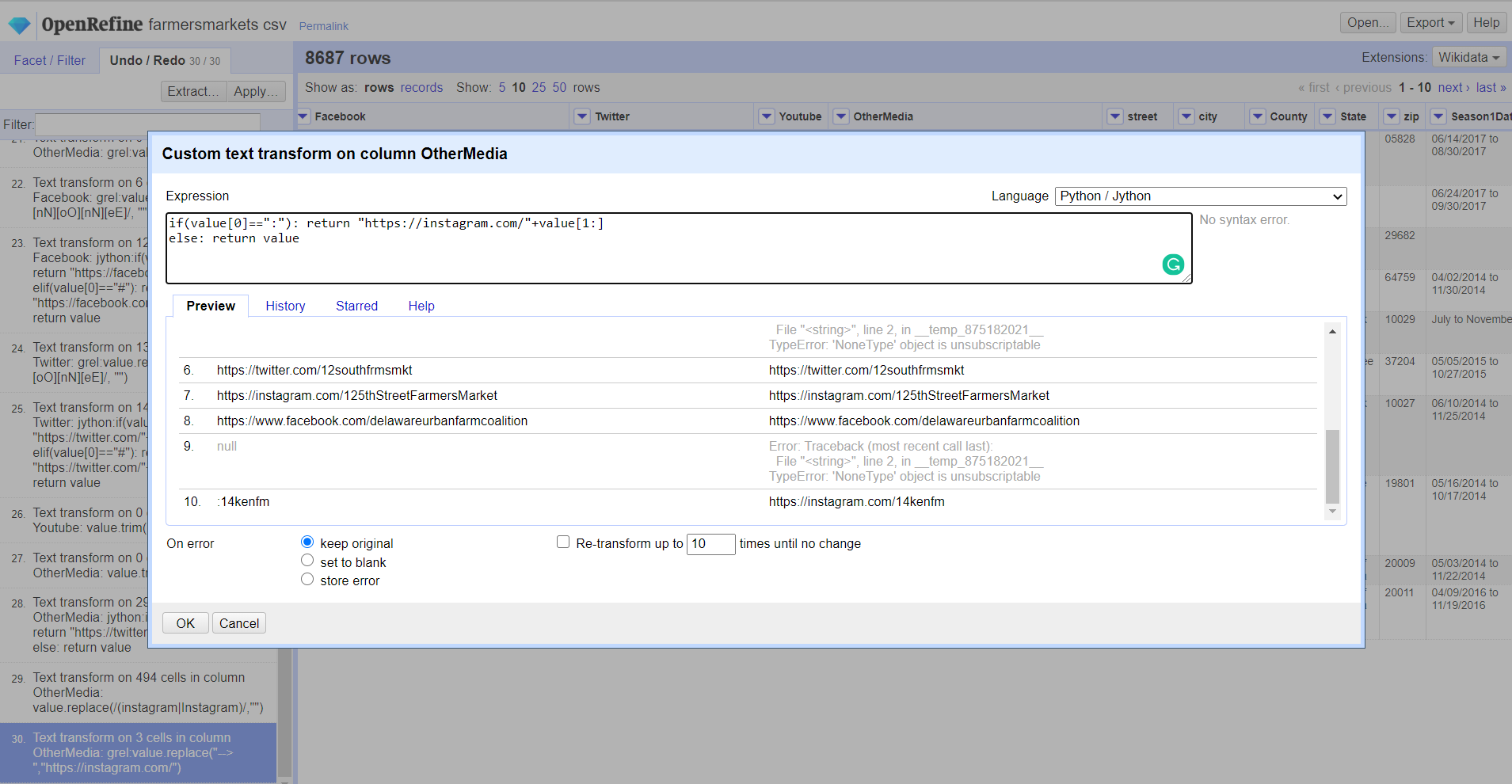


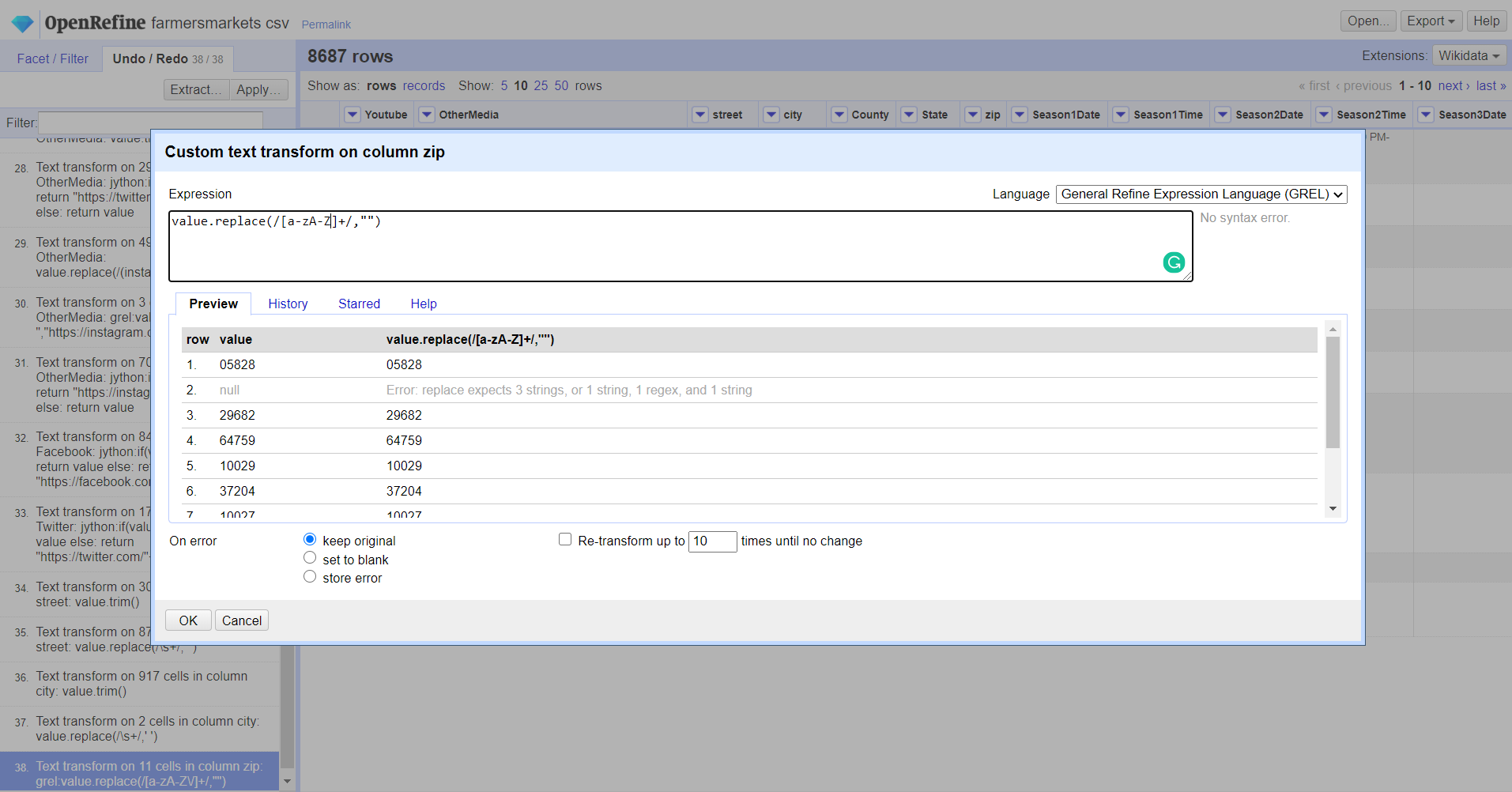












Provenance information from OpenRefine:

The complete series of steps is provided as a JSON file names schetry2\_final.json as supplementary project material.

1. **Developing a relational schema:**

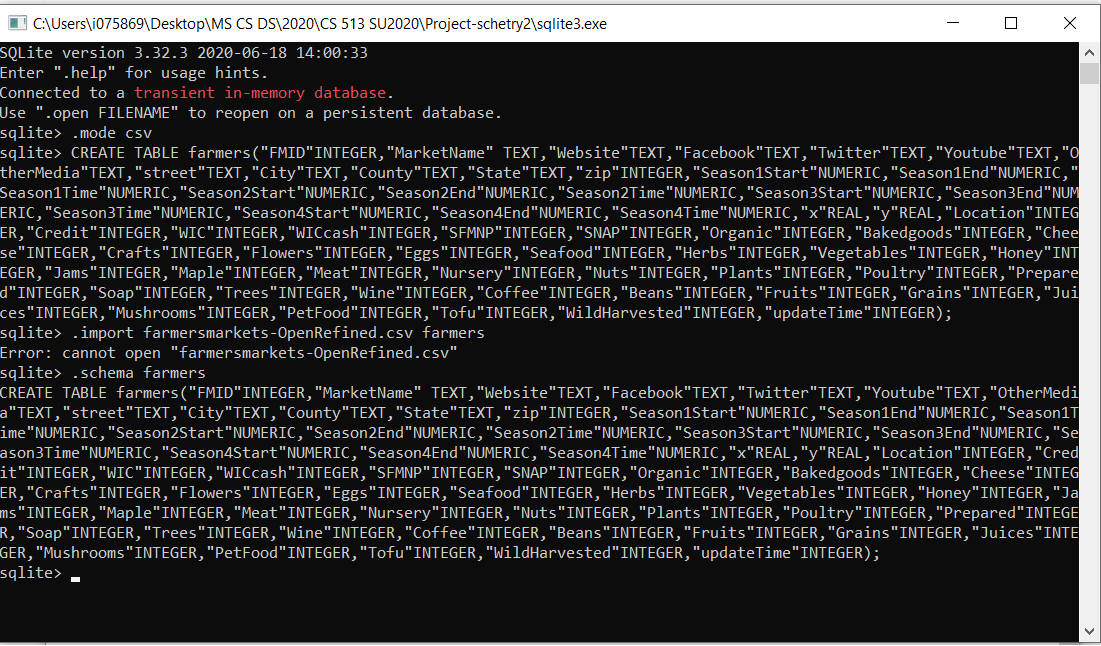
SQLite was used to create schema, load data from cleaned data from OpenRefine. SQL scripts were run to check Integrity Constraints and the violation cases were updated or removed using SQL. Finally, the SQL cleaned data was saved as csv.

Integrity Constraints identified:

1. Check for null or non-unique FMIDs
2. Check for markets which do not have ZIP codes but have X, Y co-ordinates
3. Check for markets which have X,Y co-ordinates but none of ZIP,State,City,County,Street,Social Media
4. Check for markets which do not have X,Y co-ordinates but none of ZIP,State,City,County,Street,Social Media
5. Check for duplicate social media URLs in OtherMedia field. For example Facebook URL available in both Facebook field and OtherMedia field.
6. Check for Season Start Date is older than Season End Date.
7. Check for Latitude and Longitude range values to validate.

Loading OpenRefined data into SQLite:

SQLite CLI was used for creating schema and loading the csv. SQL script was used to create schema.



Writing Queries to Check Constraints and Data Cleaning:

The list of SQL queries to check IC and update data for IC violations are provided as supplementary file.

1. Data Cleaning with Python:

There were some values in the field OtherMedia not having a valid URL format. These values were removed using Python as shown below.

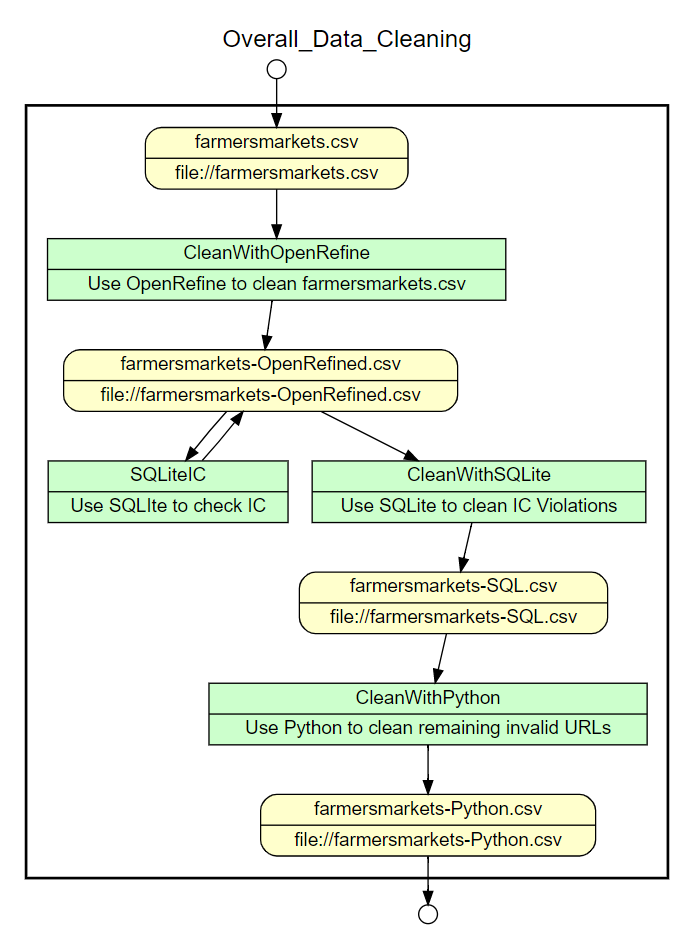
The OtherMedia values were missing domain names, which could be either facebook/twitter/Instagram. But without the domain names – the values do not add any information for locating the market. The below Python step was done to identify these non-meaningful URLs and remove them.

568 entries in OtherMedia violated correct URL and were cleaned with this operation.

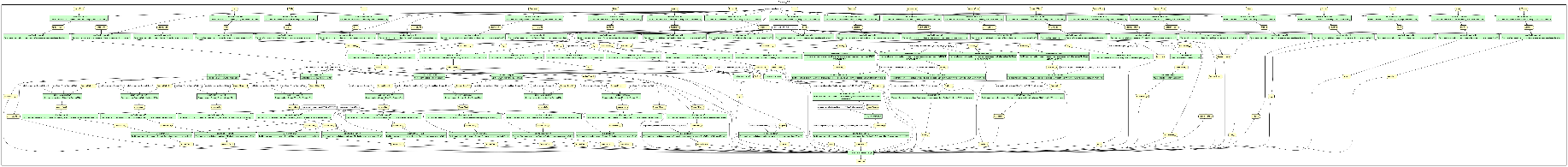


1. Creating a Workflow Model:

YesWorkFlow web interface(try.yesworkflow.org) was used for prepare the overall workflow model. Appropriate inputs and outputs from the aforementioned steps were identified and captured. The YW script and the GZ script are provided as supplementary material.



OR2YW tool was used to create the YW script. Yesworkflow web interface was used to create the OpenRefine detailed steps. The GZ file and YW script is provided as supplementary materials.



1. Further Analysis/Challenges:

Missing ZIP Codes could be replaced using Google APIs which can take state, county, city, street as input and return a JSON with ZIP code.

For entries which no ZIP codes or any location details or social media, they are not good for end user. They can be best for record keeping use.

Some values in OtherMedia had missing domain names in the URL string. This could be populated via trial and error using either facebook,twitter,youtube or instagram as probable domains.

Parsing error was encountered when converting OpenRefine JSON steps to a Yesworkflow script using OR2YW tool.

With the current cleaned data, it is now possible to build a Search and Navigation App, integrated with Map App for directions and route recommendation.