CS-513 Theory and Practice of Data Cleaning

Yanislav Shterev

The purpose of this document is to summarize the steps that were performed in order to clean and reorder the data from farmersmarkets datasets. After performing all of these operations the use cases included in the initial data assessment shall be easily to be implemented and executed.

1. **Initial data and assessment**- includes all the initial overview of the data and what are some of main issues and useful use cases that can be done based on the data. More information can be found in "1. InitialAssesment" folder "Overview & Initial Assessment.docx".

2. **OpenRefine**- Given all the data present in .csv format here are some of the columns and the transformation that were done on them. OpenRefine has been used mainly to remove bad/special characters, clustering based on similarity function, splitting the data based on some criteria and converting data to specific ISO format for consistency.

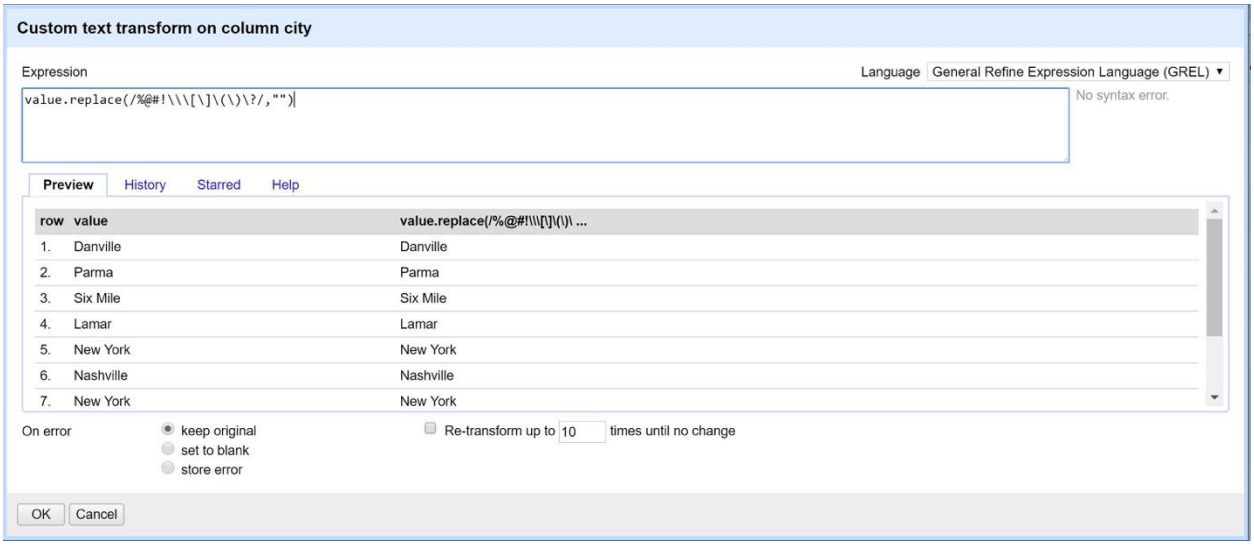
   The consolidated farmersmarkets.csv was split into 11 sub csv files which contain the same data having some relations between the different tuples. This was done to reduce the data repetitions. Now counties, cities, states etc. have their separate data csv files. On top of this the following several data cleaning operations were performed on some of the columns. The provenance information is saved in "provenance.json" file which is in the "2. OpenRefine" directory.

Transformations over **FMID**:

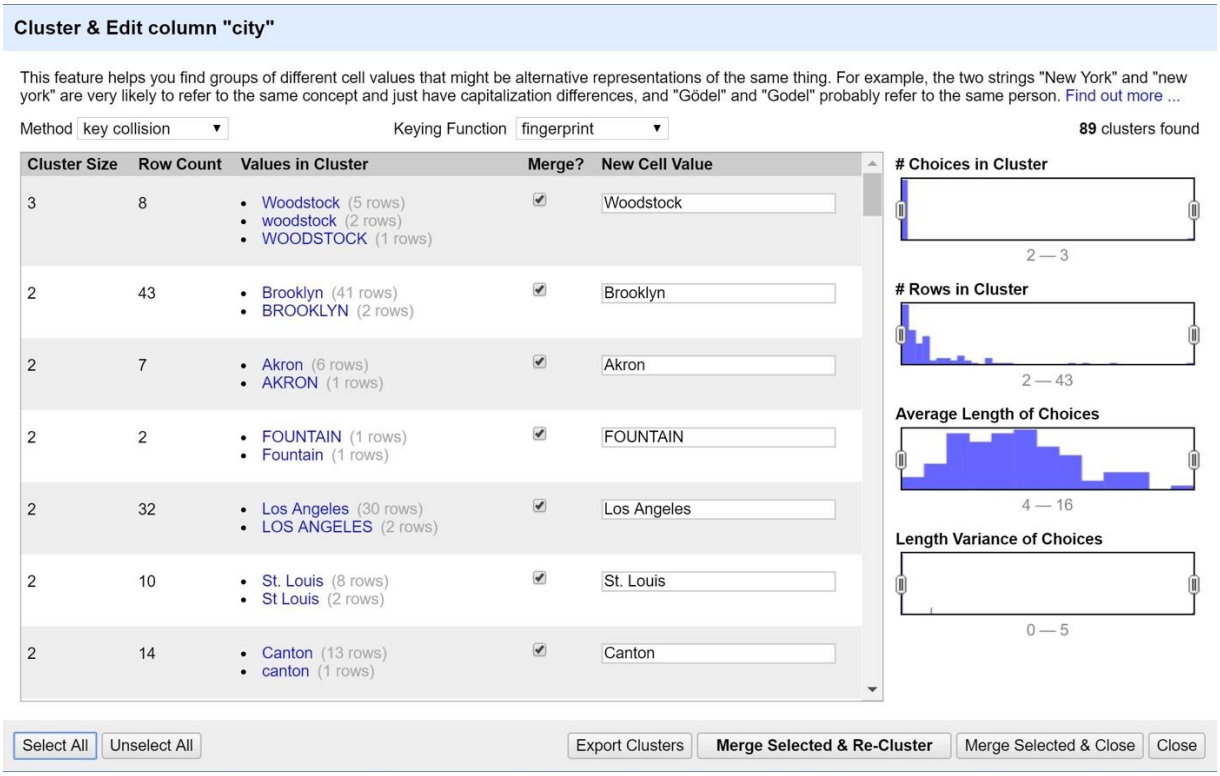
- Verify FMID contains only unique values:
  - Sort the values in FMID.
  - Blank down and verify that 0 cells were blanked down.

Transformations over **City, Country, State**:

- Remove the special characters - *% @ # ! \ [ ] ( ) ? '* using value.replace(/%@#!\\\[\]\(\)\?/,"")

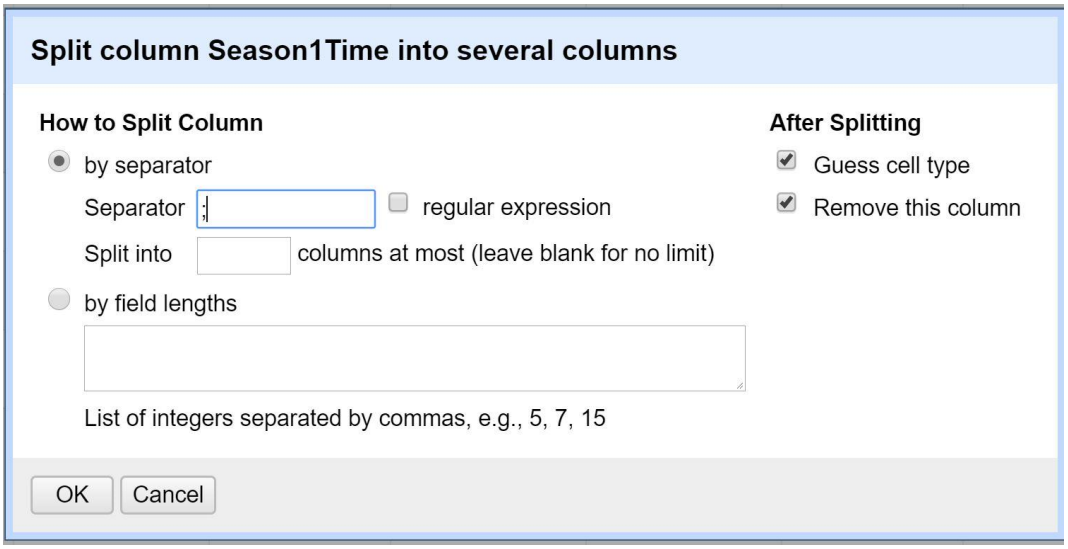


- Trim leading and trailing white spaces and collapse consecutive white spaces.
- Make a facet and perform the cluster operation using the *key-collison* method and *fingerprint* function. Merge the relevant clusters.

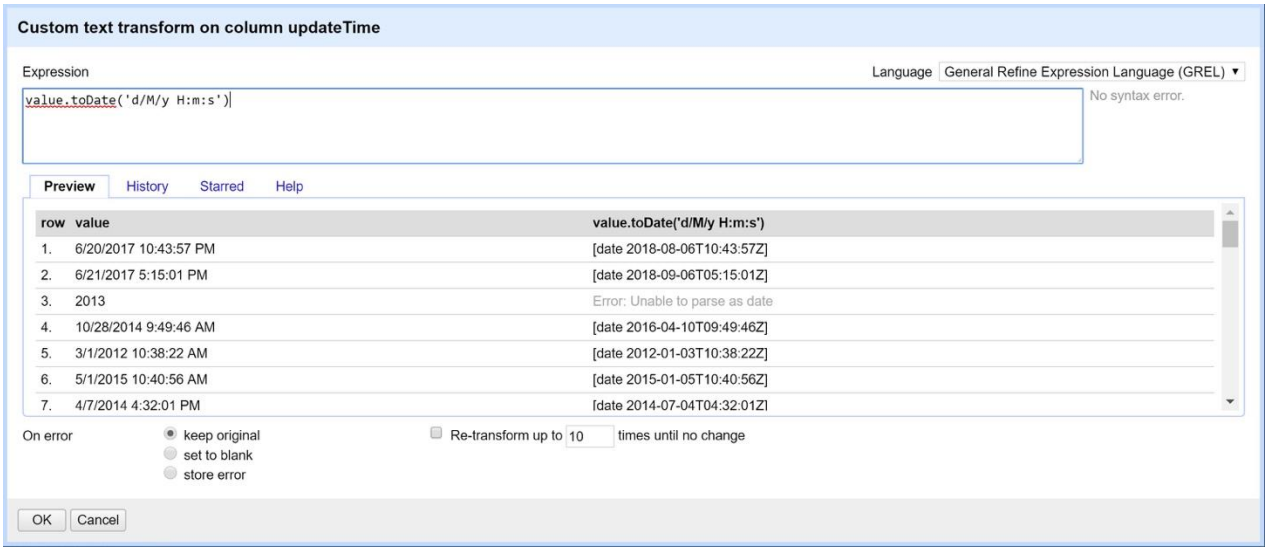


Transformations over **Season1Date**:

- Remove the special characters - *% @ # ! \ [ ] ( ) ? '* using value.replace(/%@#!\\\[\]\(\)\?/,"")
- Split the column values using * to * as separator, we get 2 columns as a result.



- Rename the first column as *Season1DateStart* and second column as *Season1DateEnd*.
- Convert the values in the column to ISO 'd/M/y' format using value.toDate('d/M/y').



- Repeat 1, 2 and 3 for *Season2Date*, *Season3Date* and *Season4Date*.

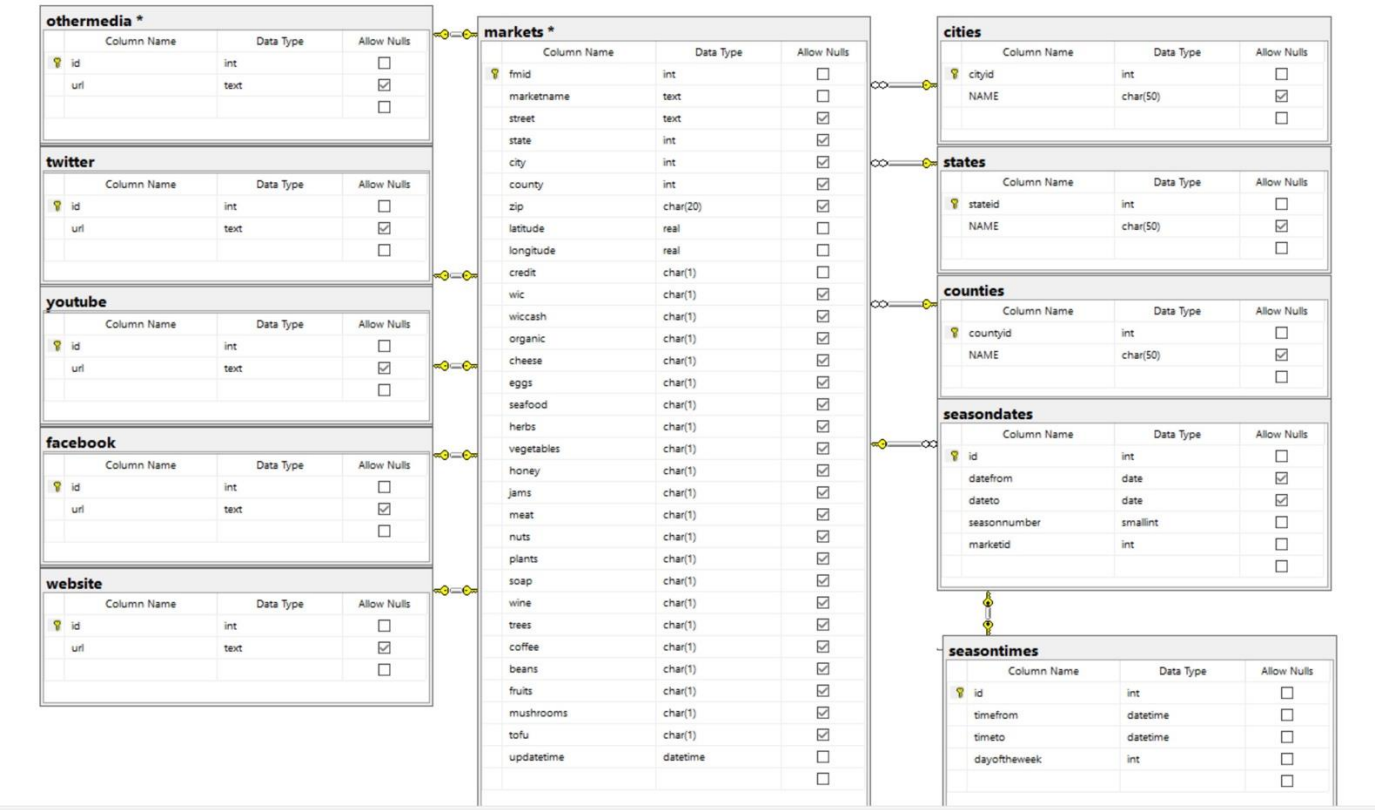
Transformations over **Season1Time**:

- Remove the special characters - *% @ # ! \ [ ] ( ) ? '* using value.replace(/%@#!\\\[\]\(\)\?/,"")
- Split the column values using *;* as separator. Each of the new columns will represent values for given day of the week.
- Repeat 1 for *Season2Time*, *Season3Time* and *Season4Time*.

Transformations over **updateTime**:

- Convert the values in the column to ISO 'd/M/y H:m:s' format using value.toDate('d/M/y H:m:s').

3. **SQL**- the following ER diagram represents the data structure after cleaning with OpenRefine and importing the data into SQLlite database. In order my database schema to satisfy the forth normal form a lot of manipulations for removing the data redundancy and improving the data reusability were done. Having the schema below it allows adaptability, expandability and high performance once the data becomes part of a real system which interacts with it. The main idea behind the division is to use the concept of dimensions and facts.



All of the scripts related to the SQL part of the project were included in sqllite-commands.txt under “3. SQL” directory. The scripts were run in a terminal to create schema, import the data from csv and export it to SQL based insert queries. “farmersmarket_schema.sql” is the file containing the plain schema implementing the relations between the tables including foreign and primary keys and other constraints like field types.

Additionally, SQL queries for integrity constrains validation were created (IC_sql_queries.sql):

- Latitude and longitude should be in the interval [0,90], [-180,180]
- There are no inactive markets. There should be at least one datetime per market in seasondates. And the datefrom and dateto values should not be empty
- Select all the markets that have seafood and vegetables and are opened in November
- There should not be two markets that are in the same location during the same time
- All the markets should have non-empty city, state and county
- No empty names for cities, states, counties

4. **YesWorkflow** – Represents the workflow of the data and how it changed during the different steps of the process. Firstly, the data root is the farmersmarkets.csv which got split into multiple files and cleaning using OpenRefine was performed over each of these files. After the cleaning the diagram represents how each of the files got imported into corresponding SQL table and constraints check was executed over all the tables to verify

whether the data has the needed quality to satisfy the use cases mentioned in the Initial Assessment document.