CS 598 Theory & Practice of Data Cleaning

# **Final Project: End-to-End Data Cleaning Workﬂow (Farmers Market)**

# **Ajay Balasubramanian**

# Overview and initial assessment

The Farmers Market Directory lists markets that feature two or more farm vendors selling agricultural products directly to customers at a common, recurrent physical location. 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.

The dataset contains 59 columns and 8687 rows of data. The first column FMID is the unique identifier for each famer’s market and is numeric. The second column is the farmer’s market name and is free text field. The next set of fields (from website to Other Media) refer to the social media urls of the market, although most of the time, these fields are not populated. The next set of fields (from street to zip) refer to the address of the location. This is again free form text, except the zip code which is supposed to have numeric fields. The next set of fields are the date/time of the seasons when the market is open. The next 2 fields refer to the coordinates of the location. The following 35 fields refer to a Boolean value of Y/N and represent the characteristic/availability of specific features within each market. The last column represents when the record was last updated.

There are quite a few data quality issues with the dataset. A number of fields have null data. URL’s of most of the social media fields are not populated and the format of the date/time is inconsistent in a large number of records. There are little variations in the text data fields that highlight the need for clustering the fields.

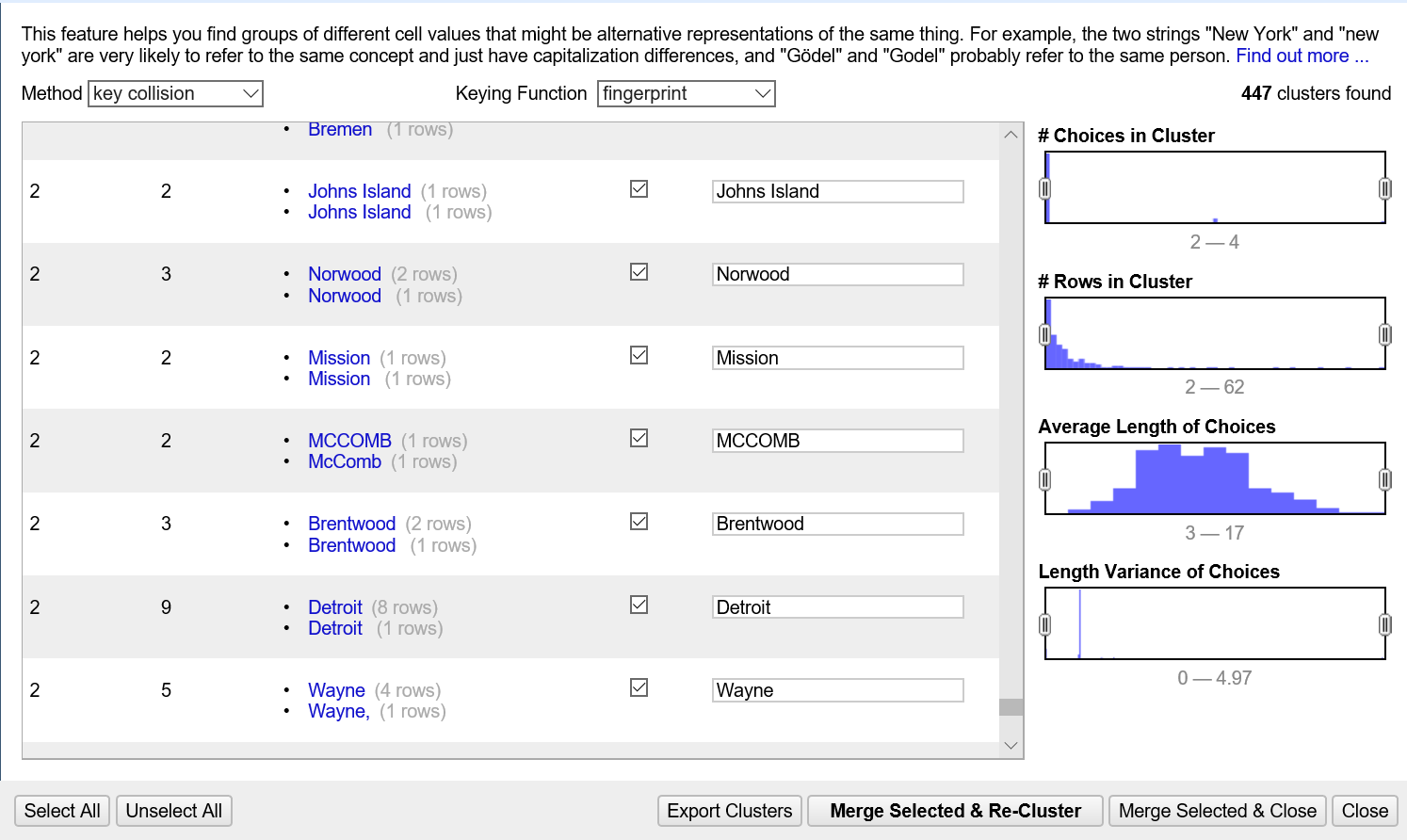
Although there are significant data consistency issues, one specific use case that can be addressed with limited data consistency issue is finding all markets that are within a specific zip code as well as the different attributes of that market – for e.g. whether it sells seafood and herbs or not. At the same time there are certain use cases for which the data is insufficient – for example, determining the customer review of each market or the current availability/pricing of individual items. One possible use case that can be answered with sufficient data cleaning and wrangling is finding out the list of markets with a certain attribute (say selling eggs) within a 5 mile radius of the user and determining whether they are currently open or not. For this we would need to clean up data of key fields like coordinates, date/time and attributes using openrefine and associating relationship between these columns using SQL.

# Data cleaning with OpenRefine

A number of data transformation and cleaning steps were performed in OpenRefine which resulted in normalization of data. Columns were chosen to be normalized based on one of the following transformations.

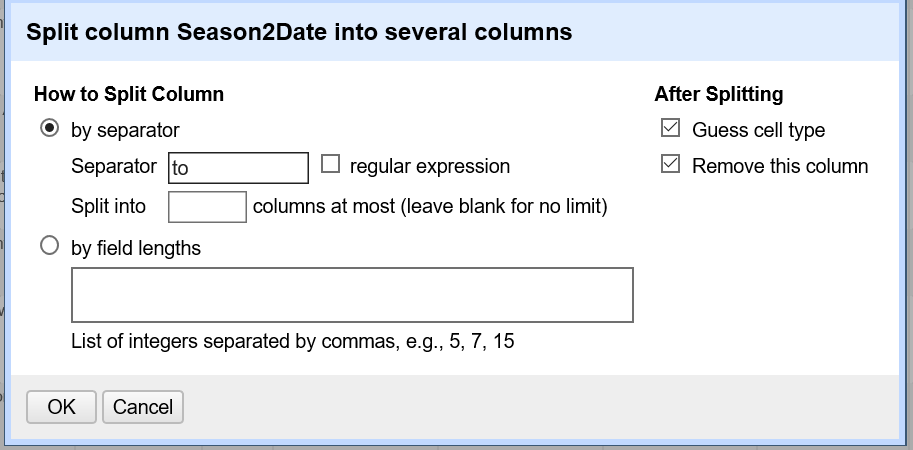
1. The following columns were clustered based on key-collision method and fingerprint function

* street
* city
* county
* state



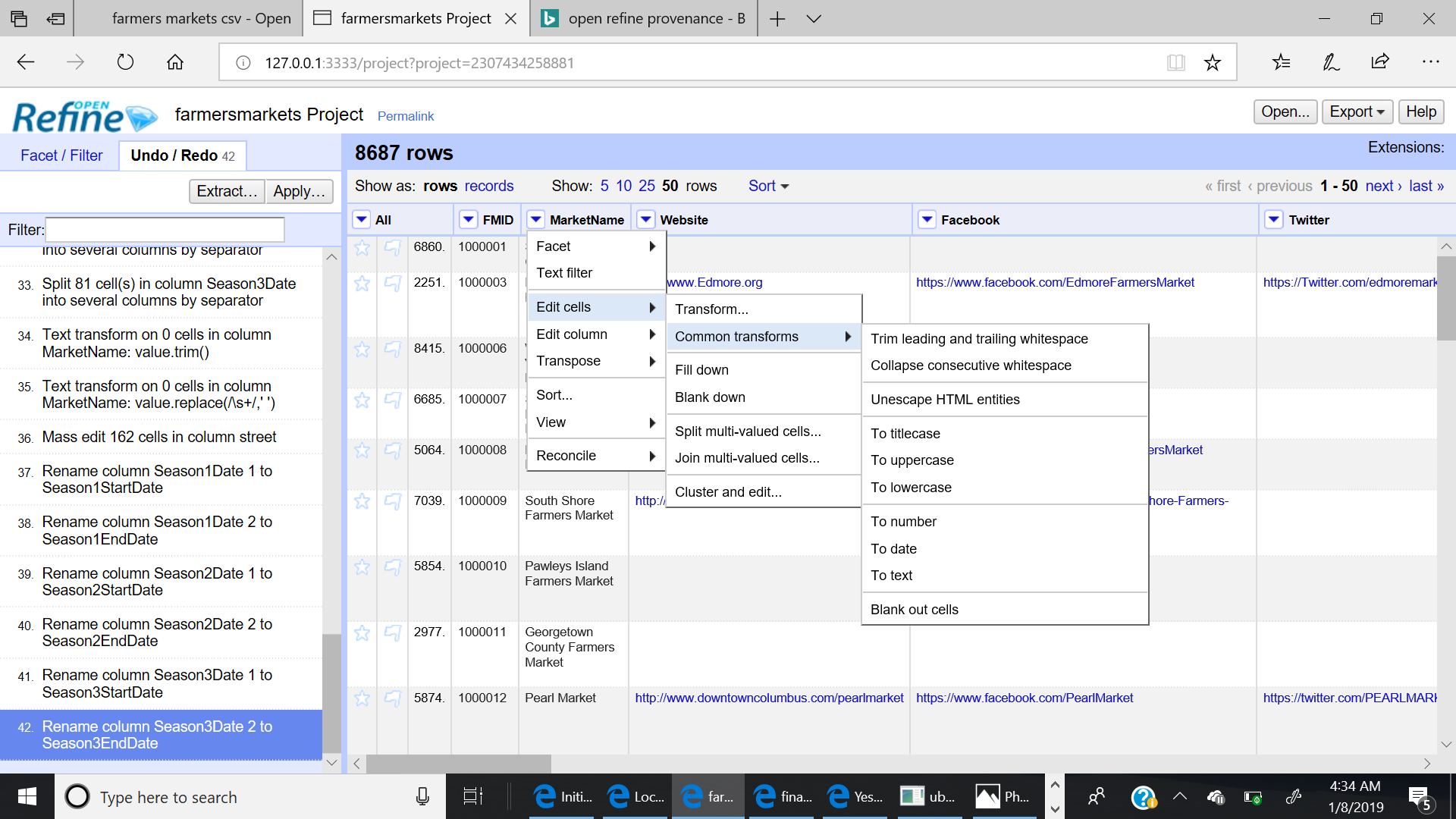
1. The following columns were split based on the delimiter “to “in order to segregate the start and end date - the split columns were named accordingly

* Season1Date
* Season2Date
* Season3Date



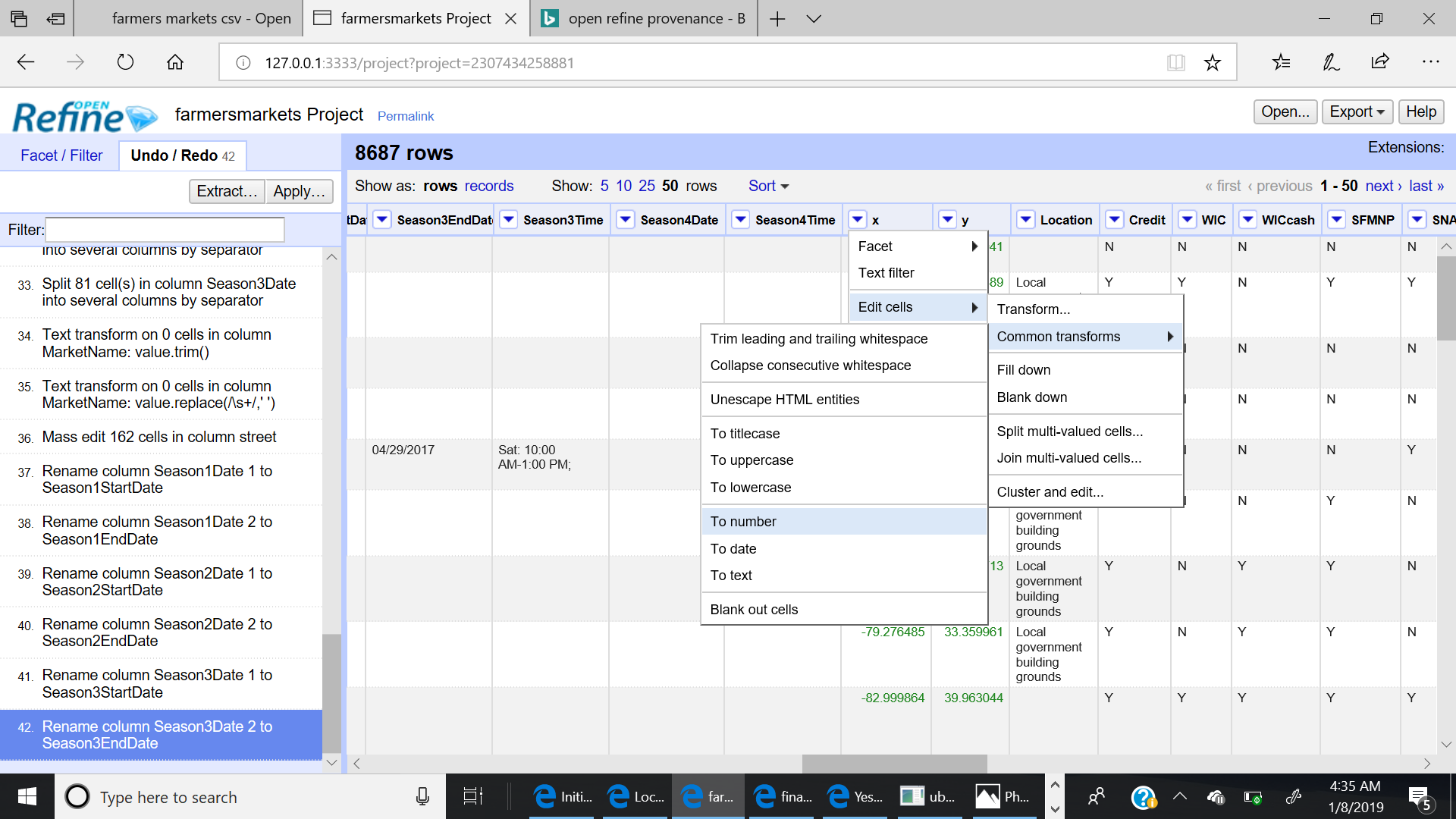
1. Leading and Trailing white spaces were cleared and consequtive whitespaces were collapsed in the following columns

* MarketName
* Facebook
* Twitter
* Website
* Youtube
* OtherMeida
* street
* city
* county
* state



1. Datatypes of the following coordinate columns were changed from text to Numeric

* x
* y



# Developing a relational database schema

The relational database schema FARMERSMARKETDB was created in SQLite Database and the output csv file from OpenRefine was loaded into it. Additionally, a few data manipulation was done to validate logical relationship between different attributes like SeasonStartDate and SeasonEndDate. The

The following integrity constraints were developed and SQL queries were created to profile the data and validate whether they satisfied these integrity constraints.

* There shall be no duplicate FMID
* There shall be no NULL FMID
* No ZIPCODE should be missing when X, Y column data are present
* Season Start date cannot be later than Season End Date

# Creating a workflow model

Workflows were created using the YesWorkflow tool to represent how data changed during the different steps of the data cleaning process. This includes both the overall Data Cleaning workflow as well as the data cleaning step within OpenRefine. Below are the visual version of the workflow.

