Theory & Practice of Data Cleaning

Introduction to OpenRefine

A First Look at OpenRefine

- Creating a New Project
- Basic Normalization
- Different Facets (text, timeline, scatterplot)
- Clustering and Mass Edits
- Operation History: Provenance

- Separate videos:
  - Installing OpenRefine
  - Advanced Operations

OpenRefine Overview

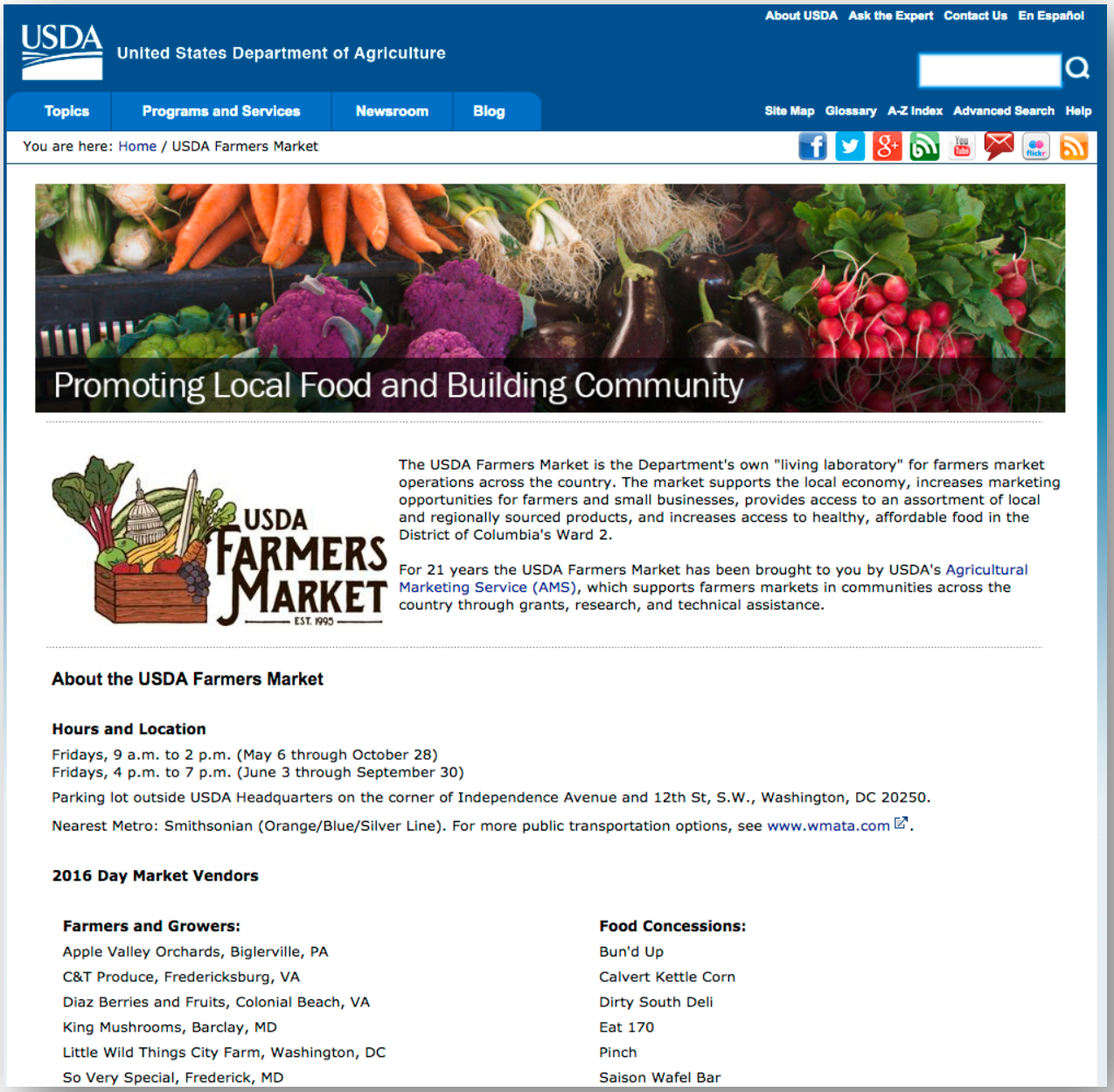
- OpenRefine is a power tool for data “wrangling”, specifically:
  - for getting an overview (exploring and “profiling”) data
  - for detecting and **cleaning** certain data errors
  - for transforming and linking data

- History:
  - Freebase Gridworks ... Google Refine … OpenRefine

Dataset Examples

- Working with two datasets:
  - USDA Directory of Farmers Markets
    - smaller, more curated (?) data
  - New York Public Library collection on historic restaurant menus
    - very “messy”, crowd-sourced data

Example: USDA Farmers Market Data



USDA United States Department of Agriculture

About USDA | Ask the Expert | Contact Us | En Español

Topics | Programs and Services | Newsroom | Blog

Site Map | Glossary | A-Z Index | Advanced Search | Help

You are here: Home / USDA Farmers Market

Promoting Local Food and Building Community

USDA FARMERS MARKET EST. 1995

The USDA Farmers Market is the Department's own "living laboratory" for farmers market operations across the country. The market supports the local economy, increases marketing opportunities for farmers and small businesses, provides access to an assortment of local and regionally sourced products, and increases access to healthy, affordable food in the District of Columbia's Ward 2.

For 21 years the USDA Farmers Market has been brought to you by USDA's Agricultural Marketing Service (AMS), which supports farmers markets in communities across the country through grants, research, and technical assistance.

About the USDA Farmers Market

Hours and Location

Fridays, 9 a.m. to 2 p.m. (May 6 through October 28)
Fridays, 4 p.m. to 7 p.m. (June 3 through September 30)

Parking lot outside USDA Headquarters on the corner of Independence Avenue and 12th St, S.W., Washington, DC 20250.

Nearest Metro: Smithsonian (Orange/Blue/Silver Line). For more public transportation options, see www.wmata.com.

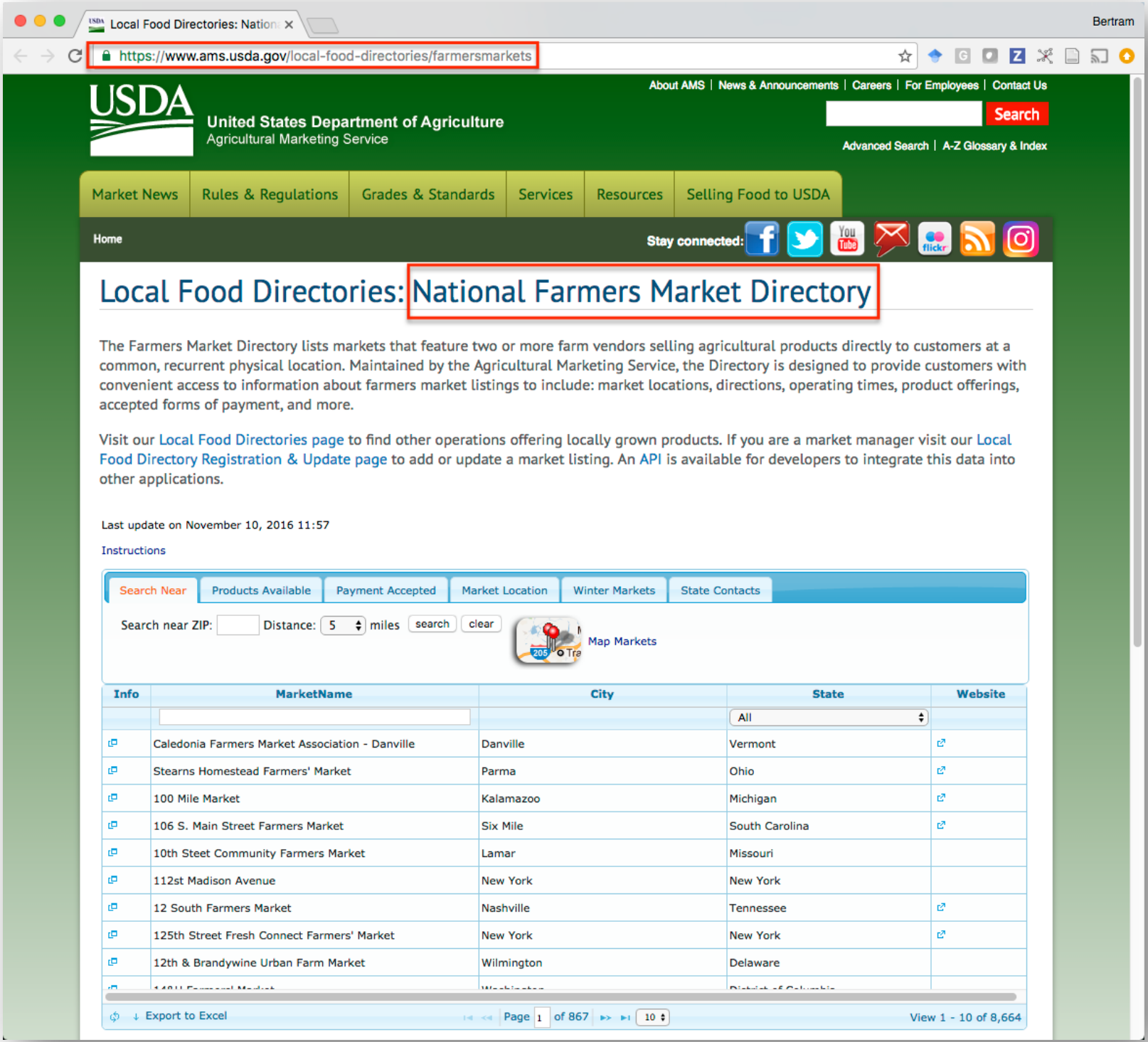
2016 Day Market Vendors

Farmers and Growers:

Apple Valley Orchards, Biglerville, PA
C&T Produce, Fredericksburg, VA
Diaz Berries and Fruits, Colonial Beach, VA
King Mushrooms, Barclay, MD
Little Wild Things City Farm, Washington, DC
So Very Special, Frederick, MD

Food Concessions:

Bun'd Up
Calvert Kettle Corn
Dirty South Deli
Eat 170
Pinch
Saison Wafel Bar


Local Food Directories: Nation
Bertram
https://www.ams.usda.gov/local-food-directories/farmersmarkets
About AMS | News & Announcements | Careers | For Employees | Contact Us
USDA
United States Department of Agriculture
Agricultural Marketing Service
Search
Advanced Search | A-Z Glossary & Index
Market News
Rules & Regulations
Grades & Standards
Services
Resources
Selling Food to USDA
Home
Stay connected:
Local Food Directories: National Farmers Market Directory
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.
Visit our Local Food Directories page to find other operations offering locally grown products. If you are a market manager visit our Local Food Directory Registration & Update page to add or update a market listing. An API is available for developers to integrate this data into other applications.
Last update on November 10, 2016 11:57
Instructions
Search Near
Products Available
Payment Accepted
Market Location
Winter Markets
State Contacts
Search near ZIP:
Distance: 5 miles
search
clear
Map Markets
Info
MarketName
City
State
Website
All
Caledonia Farmers Market Association - Danville
Danville
Vermont
Stearns Homestead Farmers' Market
Parma
Ohio
100 Mile Market
Kalamazoo
Michigan
106 S. Main Street Farmers Market
Six Mile
South Carolina
10th Steet Community Farmers Market
Lamar
Missouri
112st Madison Avenue
New York
New York
12 South Farmers Market
Nashville
Tennessee
125th Street Fresh Connect Farmers' Market
New York
New York
12th & Brandywine Urban Farm Market
Wilmington
Delaware
Export to Excel
Page 1 of 867
10
View 1 - 10 of 8,664

Local Food Directories: National Farmers Market Directory

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.

Visit our Local Food Directories page to find other operations offering locally grown products. If you are a market manager visit our Local Food Directory Registration & Update page to add or update a market listing. An API is available for developers to integrate this data into other applications.

Last update on November 10, 2016 11:57

Instructions

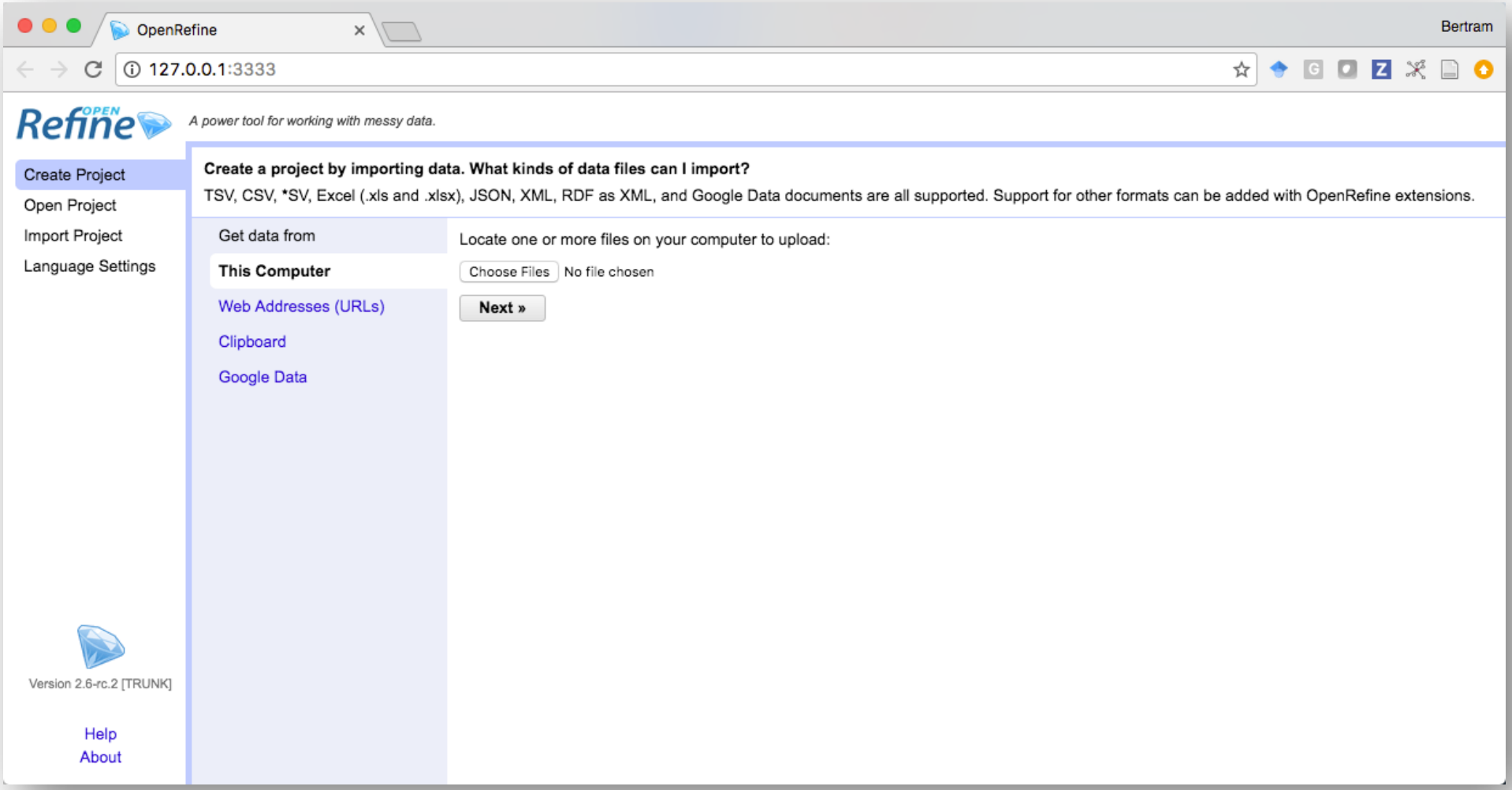
Search Near | Products Available | Payment Accepted | Market Location | Winter Markets | State Contacts

Search near ZIP: Distance: 5 miles search clear Map Markets

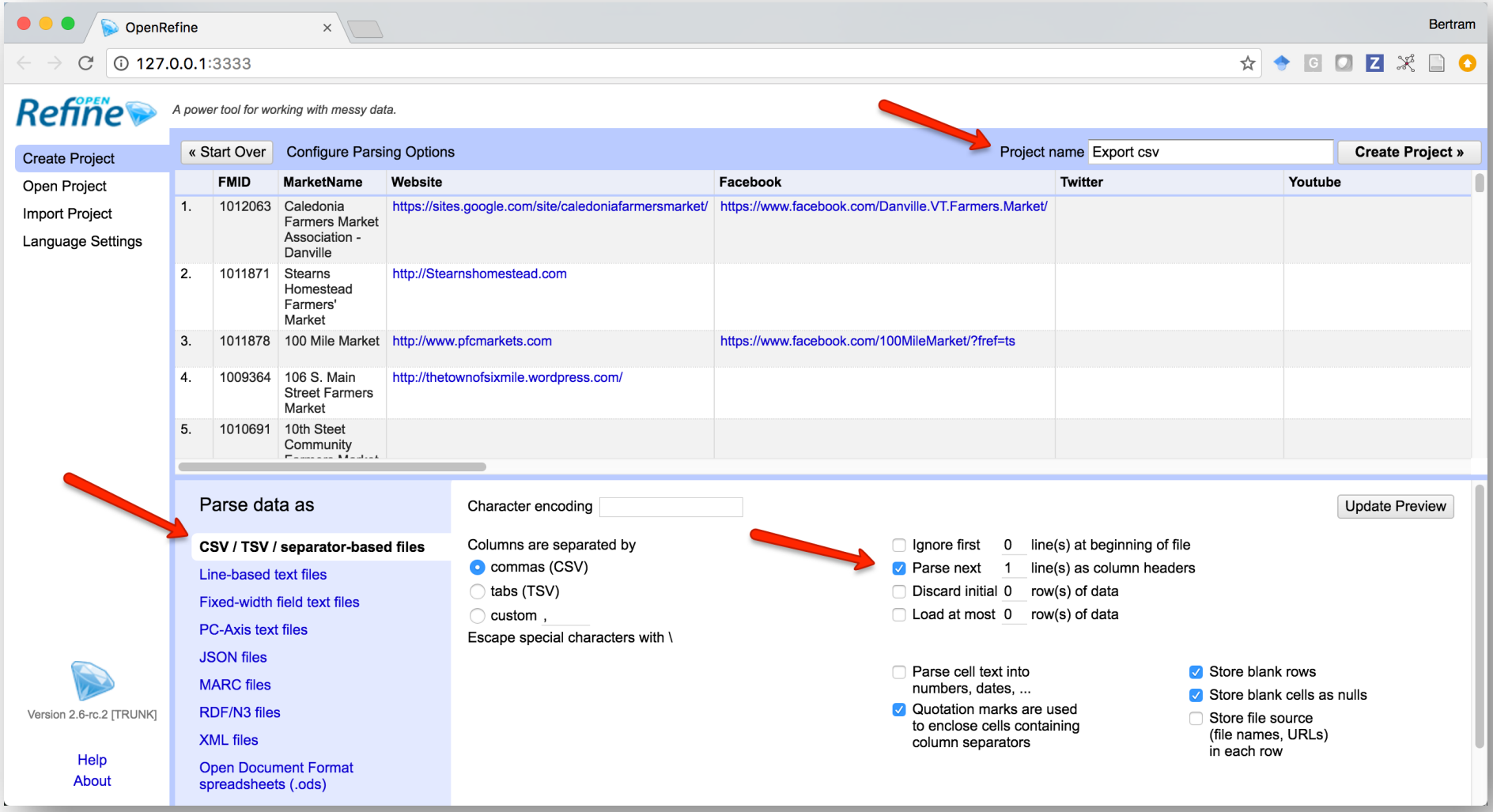
Info	MarketName	City	State	Website
			All	
	Caledonia Farmers Market Association - Danville	Danville	Vermont	
	Stearns Homestead Farmers' Market	Parma	Ohio	
	100 Mile Market	Kalamazoo	Michigan	
	106 S. Main Street Farmers Market	Six Mile	South Carolina	
	10th Steet Community Farmers Market	Lamar	Missouri	
	112st Madison Avenue	New York	New York	
	12 South Farmers Market	Nashville	Tennessee	
	125th Street Fresh Connect Farmers' Market	New York	New York	
	12th & Brandywine Urban Farm Market	Wilmington	Delaware	

Export to Excel Page 1 of 867 10 View 1 - 10 of 8,664

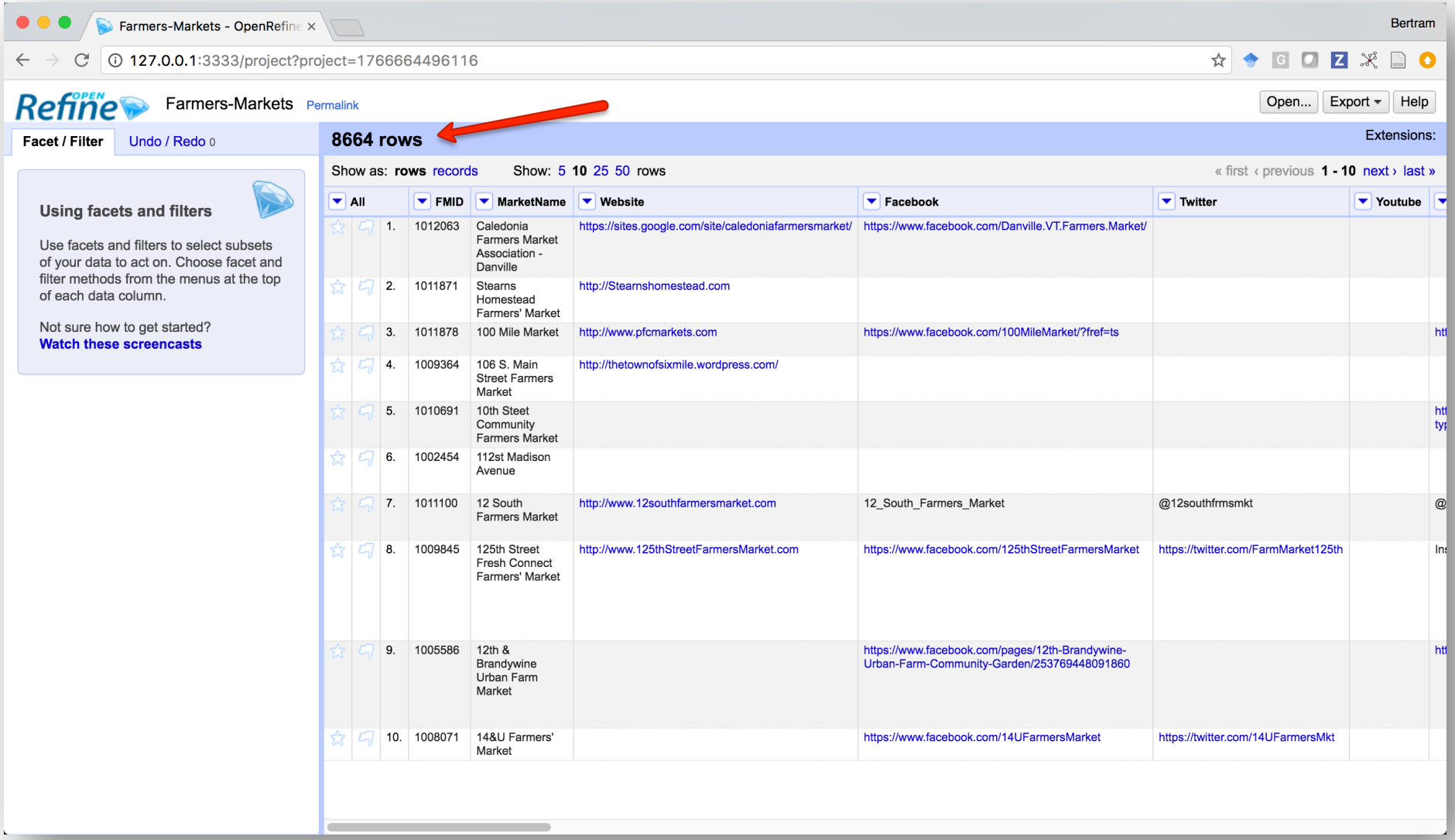
OpenRefine: Create Project



Importing Data …

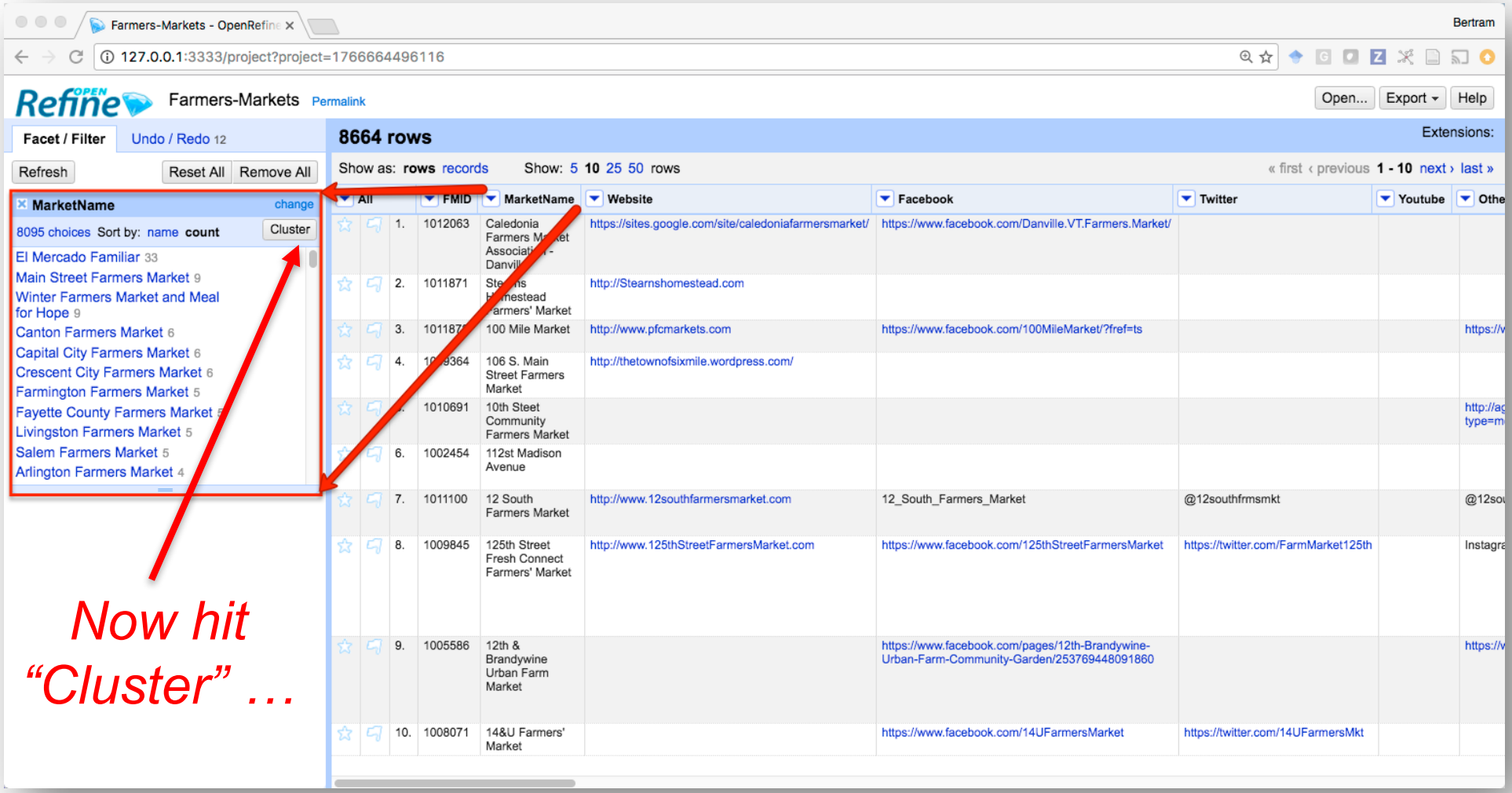


Voilà! 8664 rows imported …

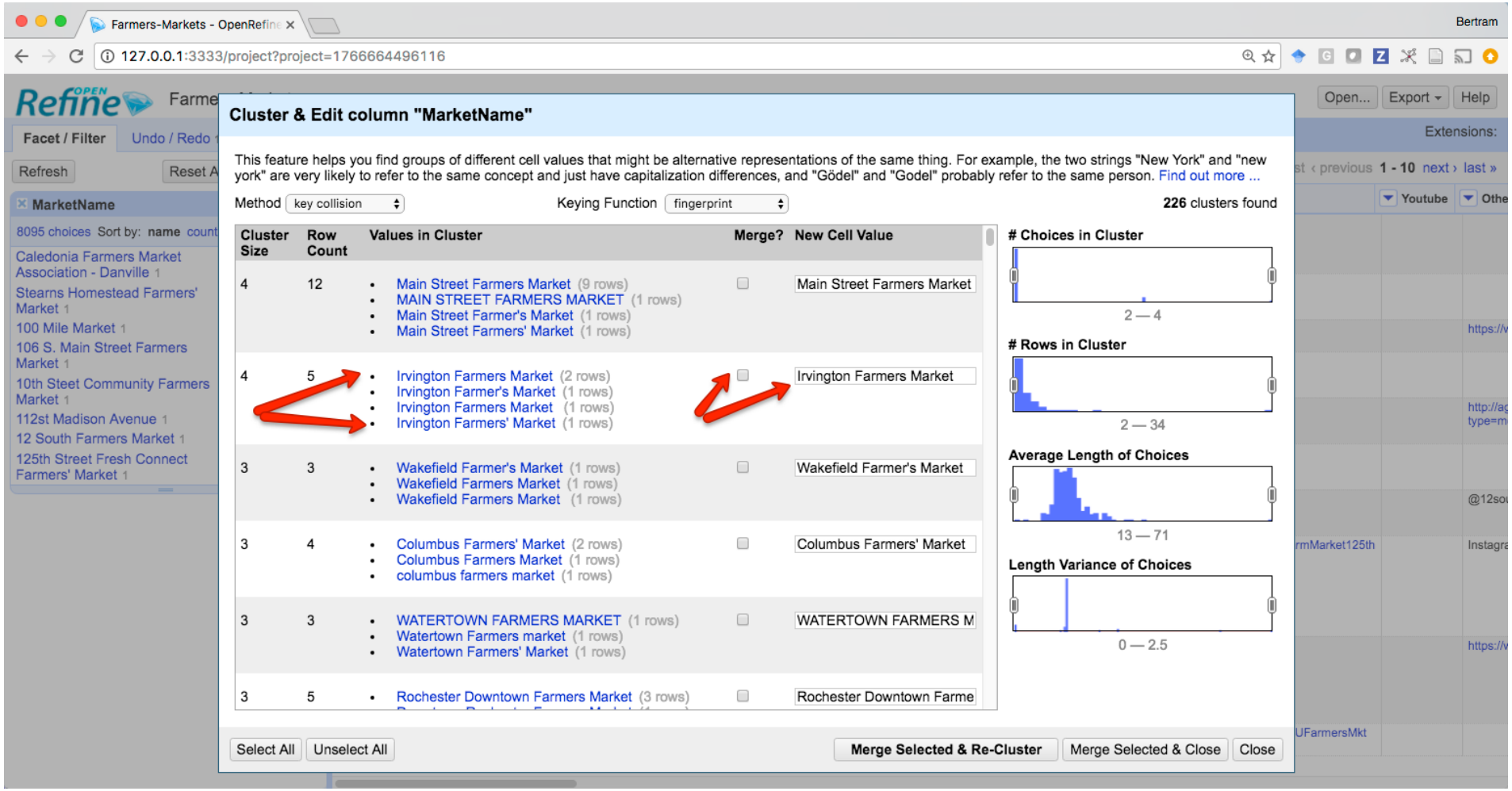


All		FMID	MarketName	Website	Facebook	Twitter	Youtube
1.	1012063	Caledonia Farmers Market Association - Danville	https://sites.google.com/site/caledoniafarmersmarket/	https://www.facebook.com/Danville.VT.Farmers.Market/			
2.	1011871	Stearns Homestead Farmers' Market	http://Stearnshomestead.com				
3.	1011878	100 Mile Market	http://www.pfcmarkets.com	https://www.facebook.com/100MileMarket/?fref=ts			
4.	1009364	106 S. Main Street Farmers Market	http://thetownofsixmile.wordpress.com/				
5.	1010691	10th Steet Community Farmers Market					
6.	1002454	112st Madison Avenue					
7.	1011100	12 South Farmers Market	http://www.12southfarmersmarket.com	12_South_Farmers_Market	@12southfrmsmkt		
8.	1009845	125th Street Fresh Connect Farmers' Market	http://www.125thStreetFarmersMarket.com	https://www.facebook.com/125thStreetFarmersMarket	https://twitter.com/FarmMarket125th		
9.	1005586	12th & Brandywine Urban Farm Market		https://www.facebook.com/pages/12th-Brandywine-Urban-Farm-Community-Garden/253769448091860			
10.	1008071	14&U Farmers' Market		https://www.facebook.com/14UFarmersMarket	https://twitter.com/14UFarmersMkt		

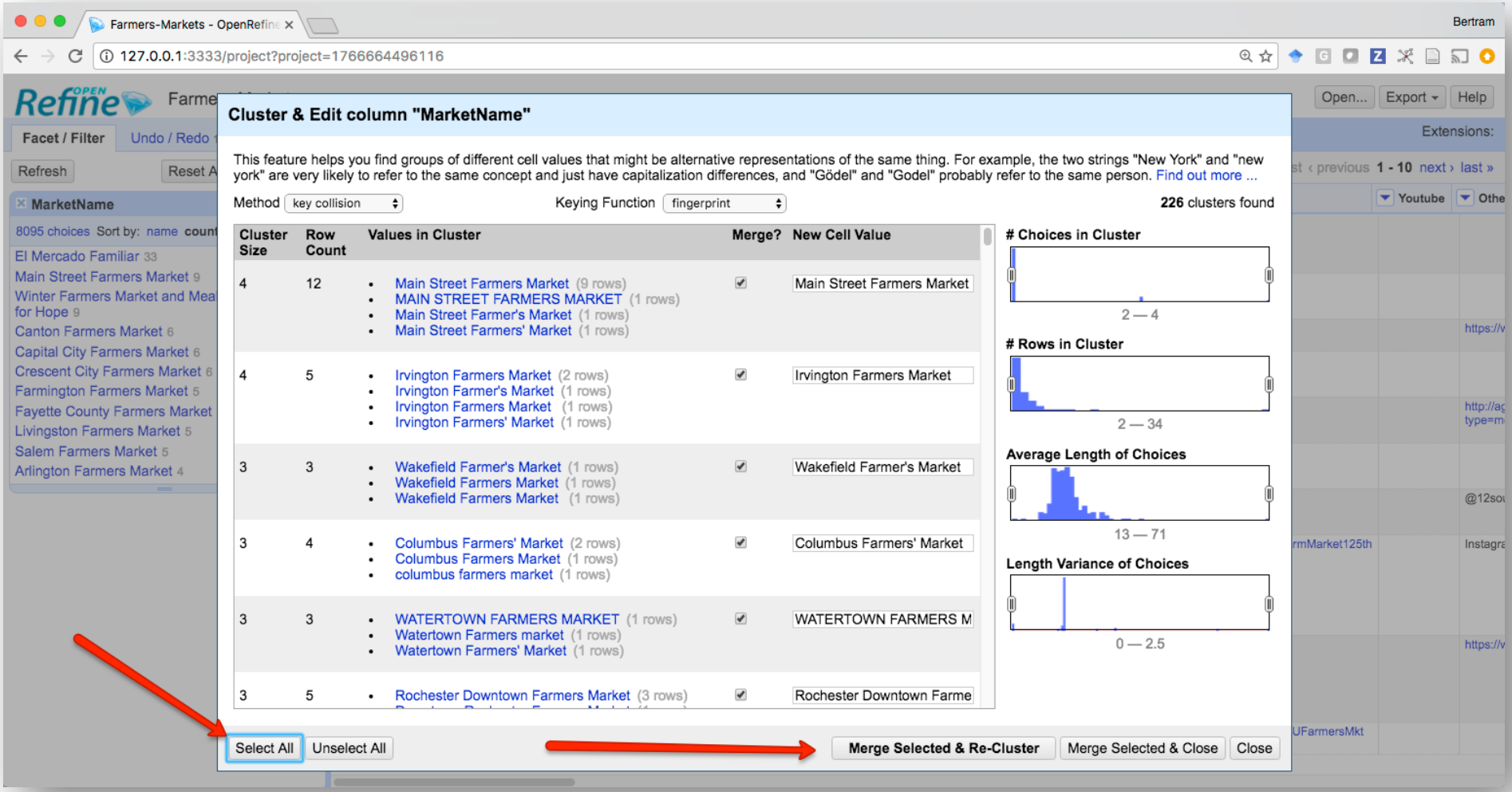
The **Text Facet** “workhorse” …



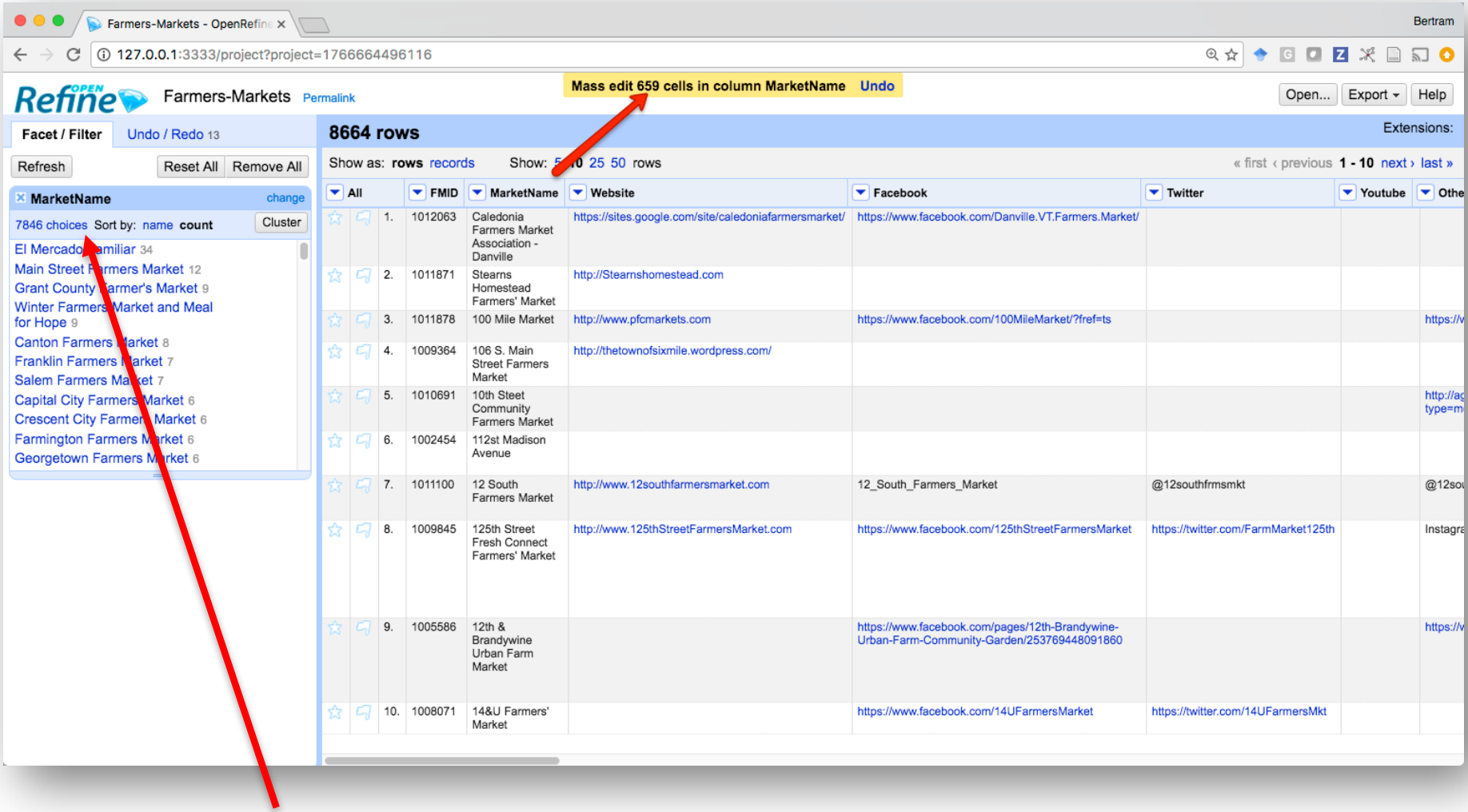
… and the magic happens!



… select some (all!?) clusters and merge …

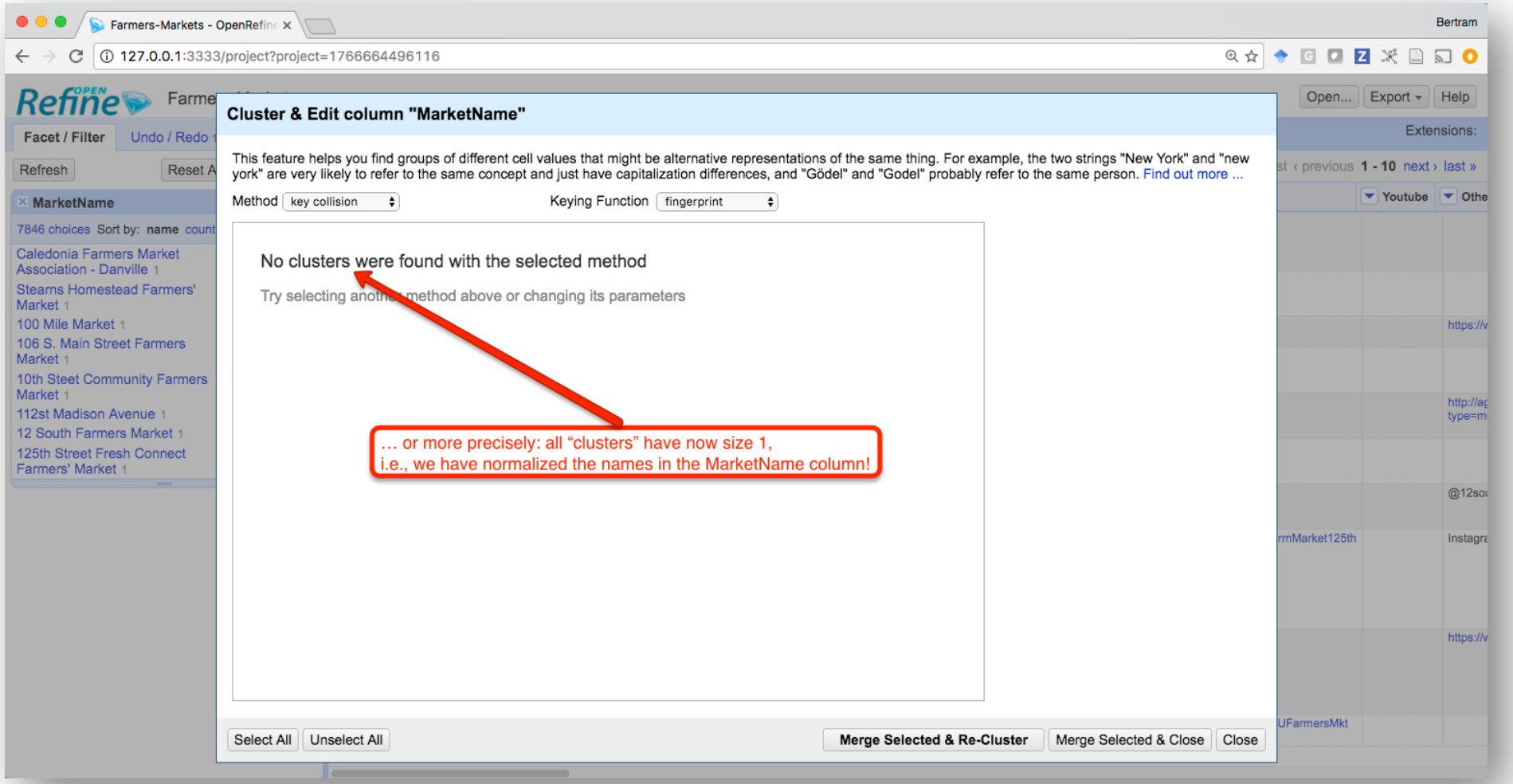


… resulting in a **mass edit** …

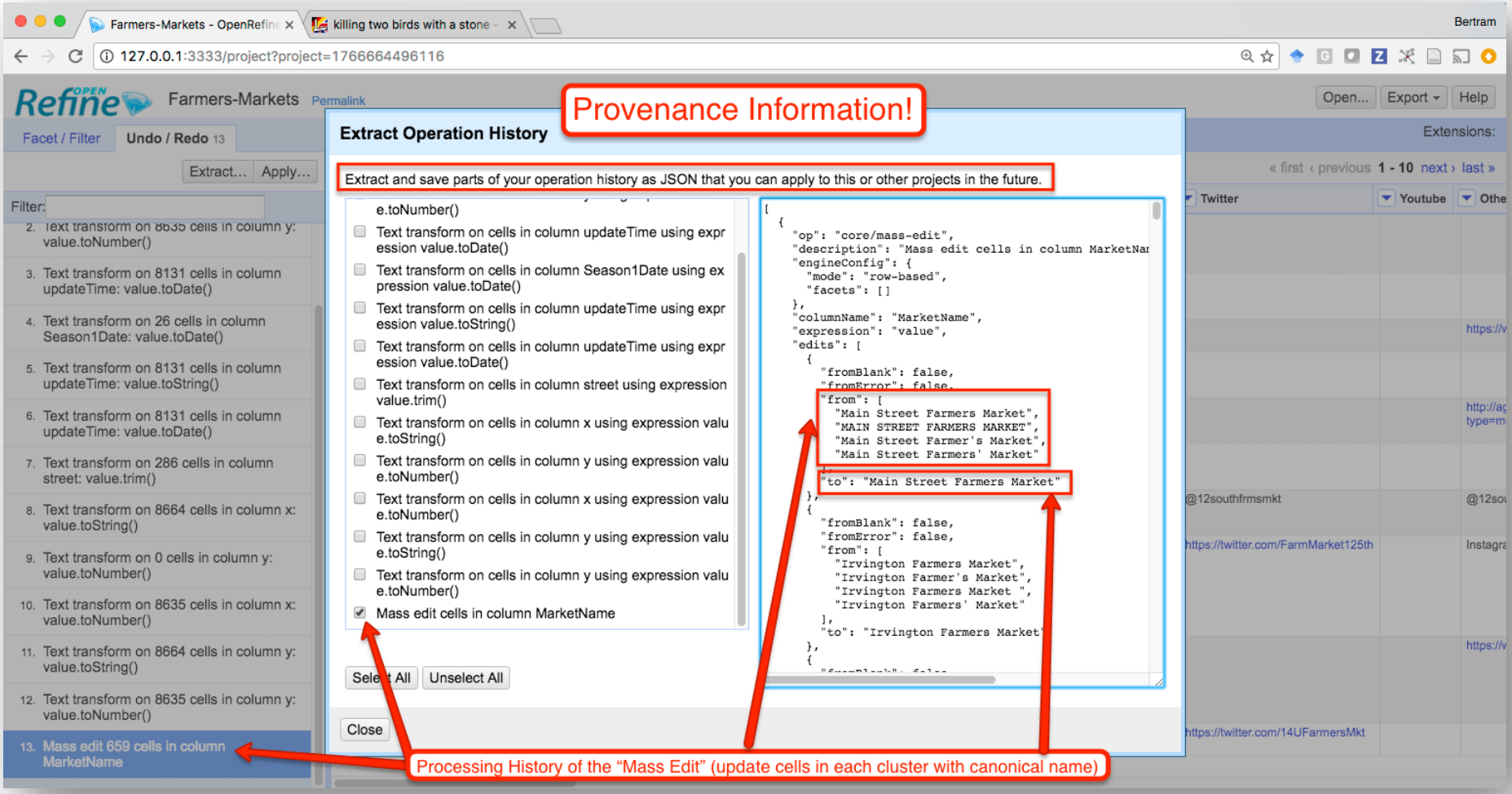


.. also reduced the choices from 8095 to 7846…

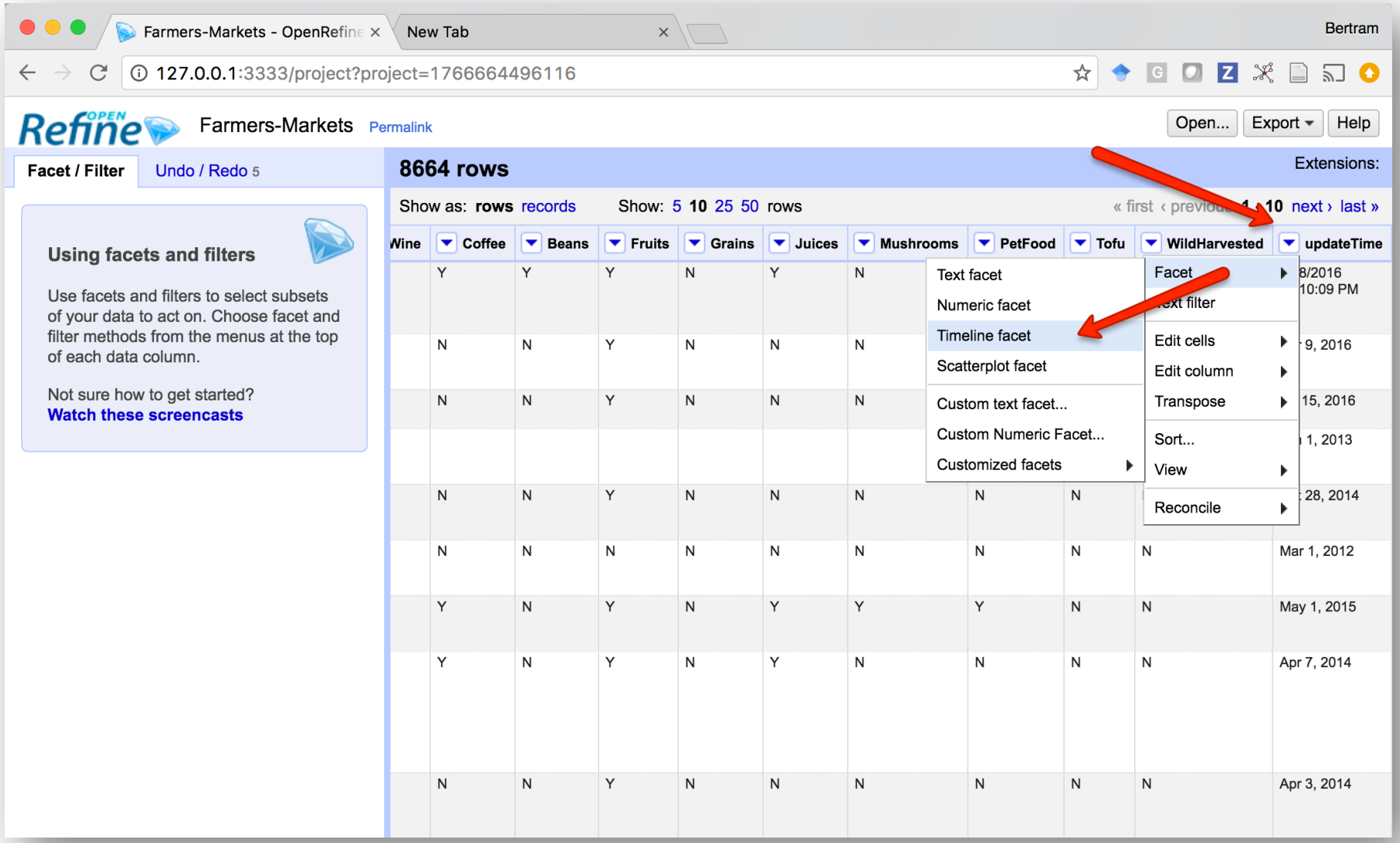
… (in this case): Done with **Normalization** of MarketName column



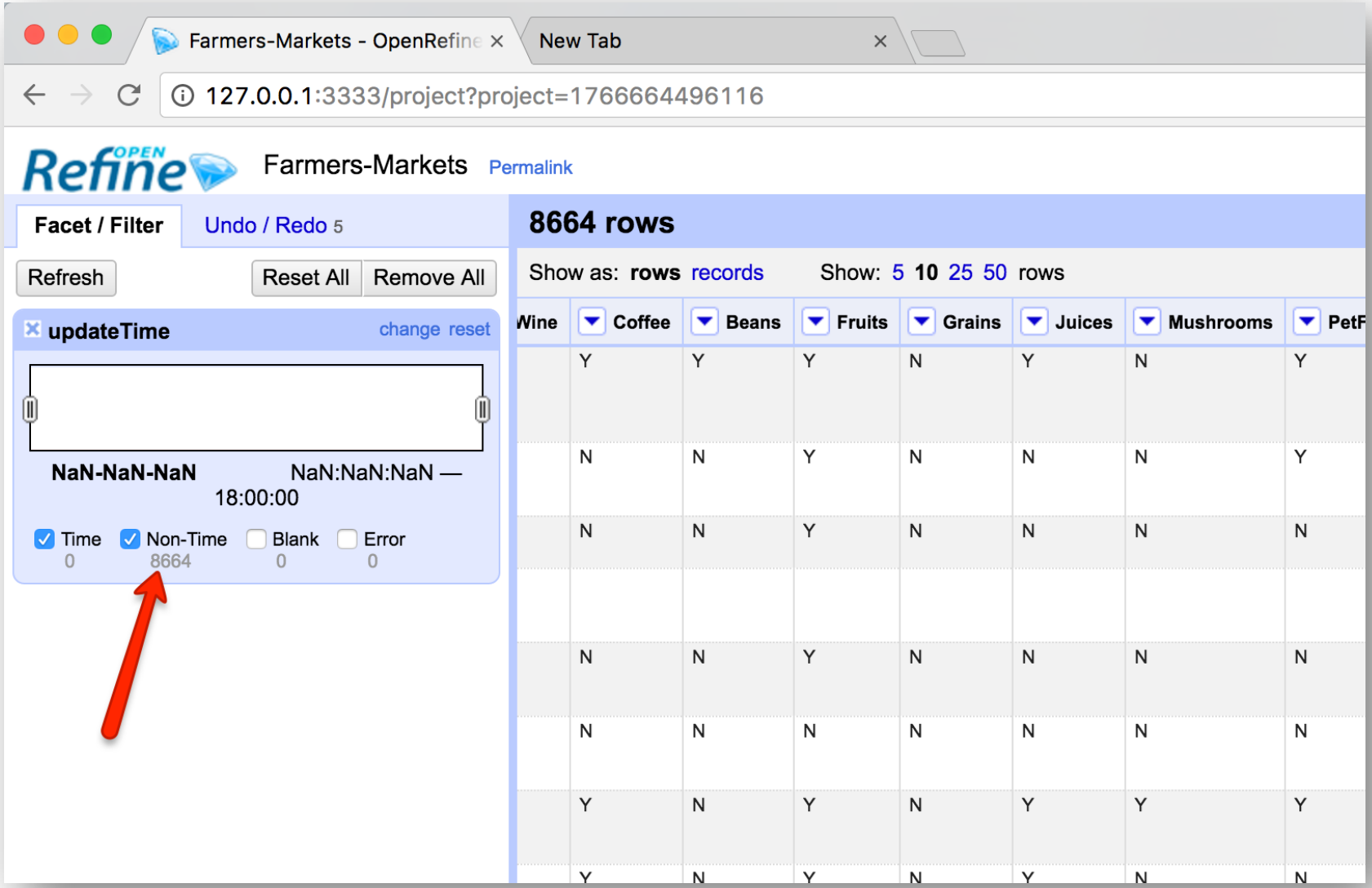
Undo/Redo: **Operation History** (***Provenance***)



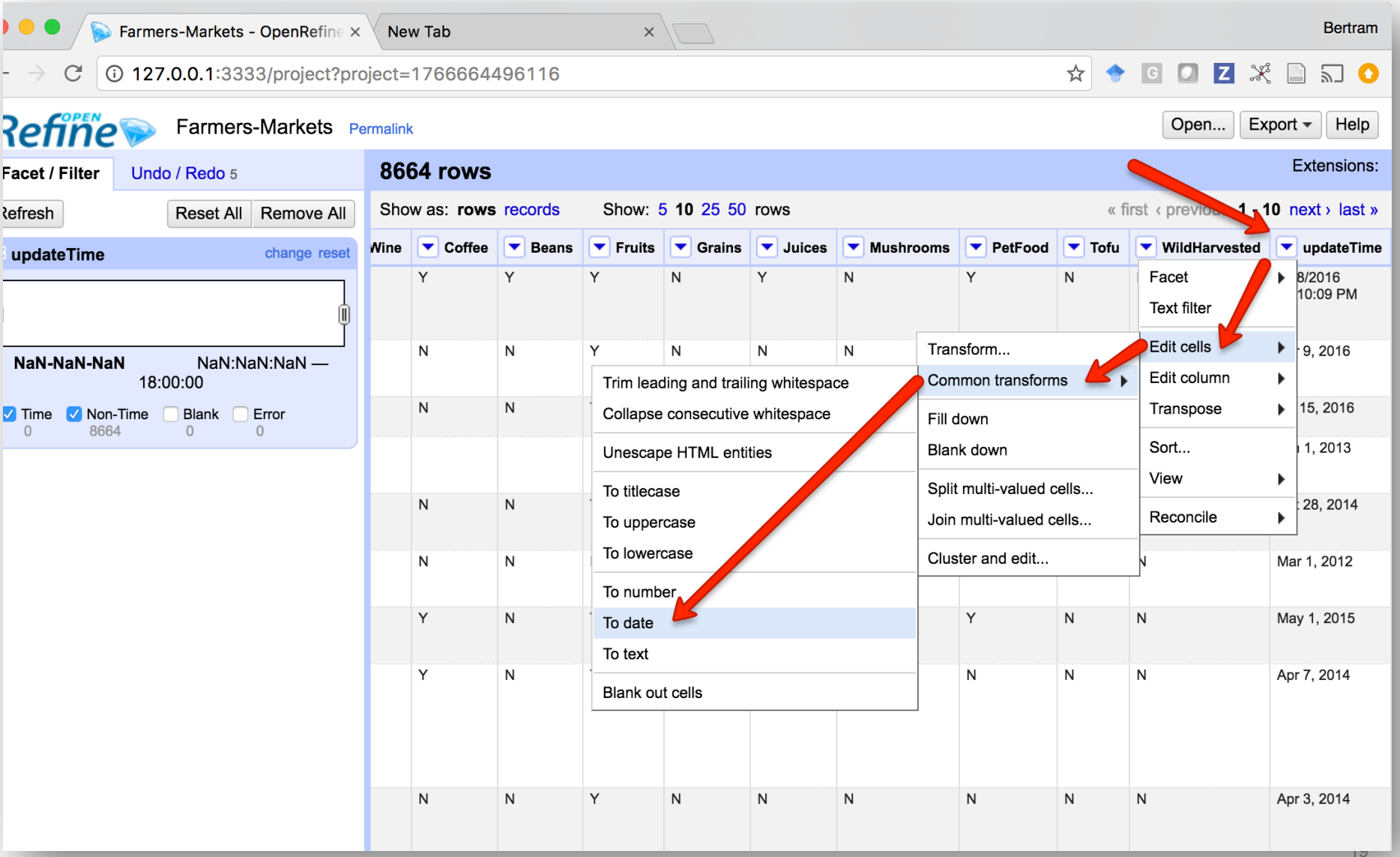
More Data Profiling: **Timeline Facet**



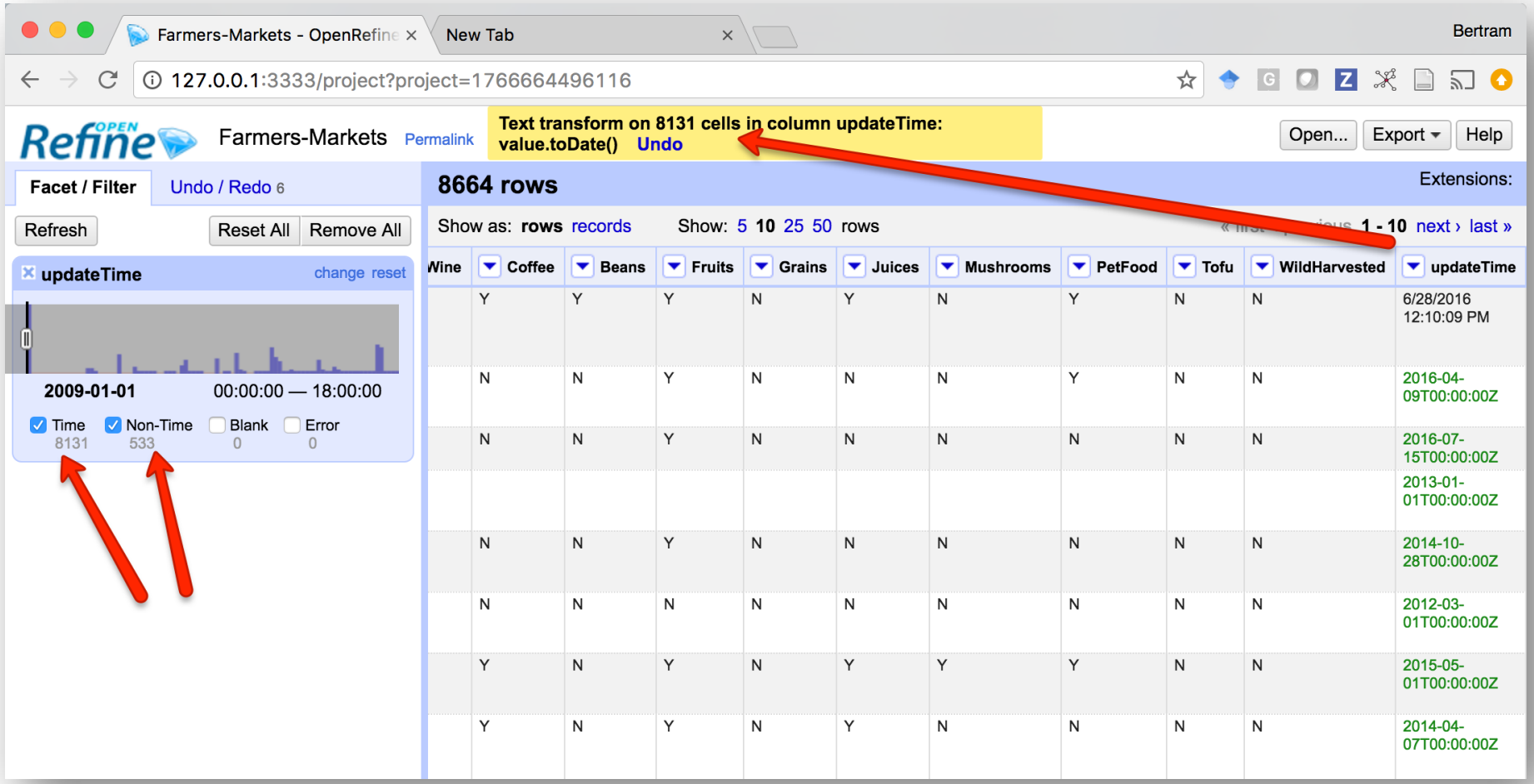
Timeline facet: hmm ... not working!?



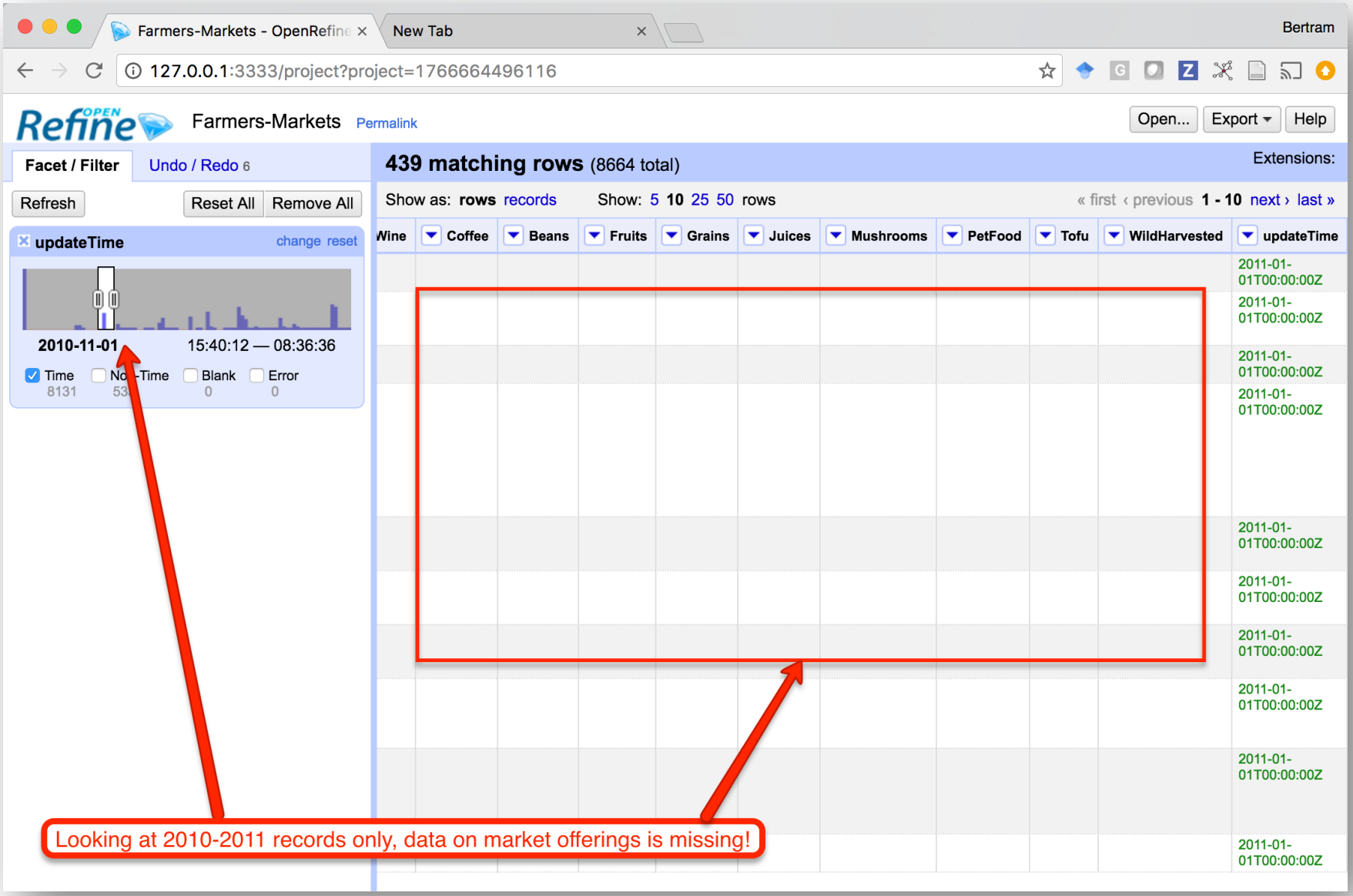
Converting from **String** to **Date!**



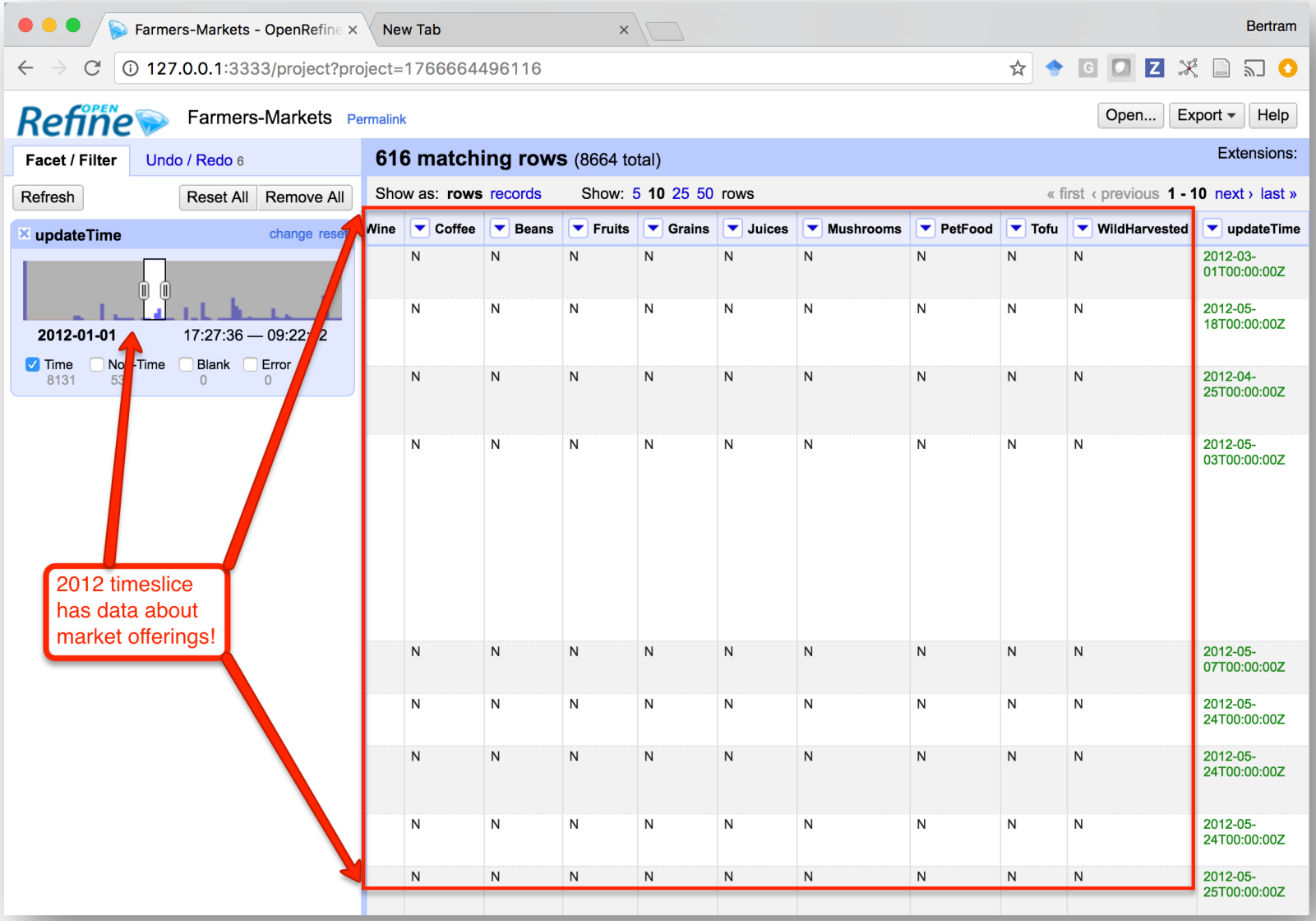
.. now we're in business!



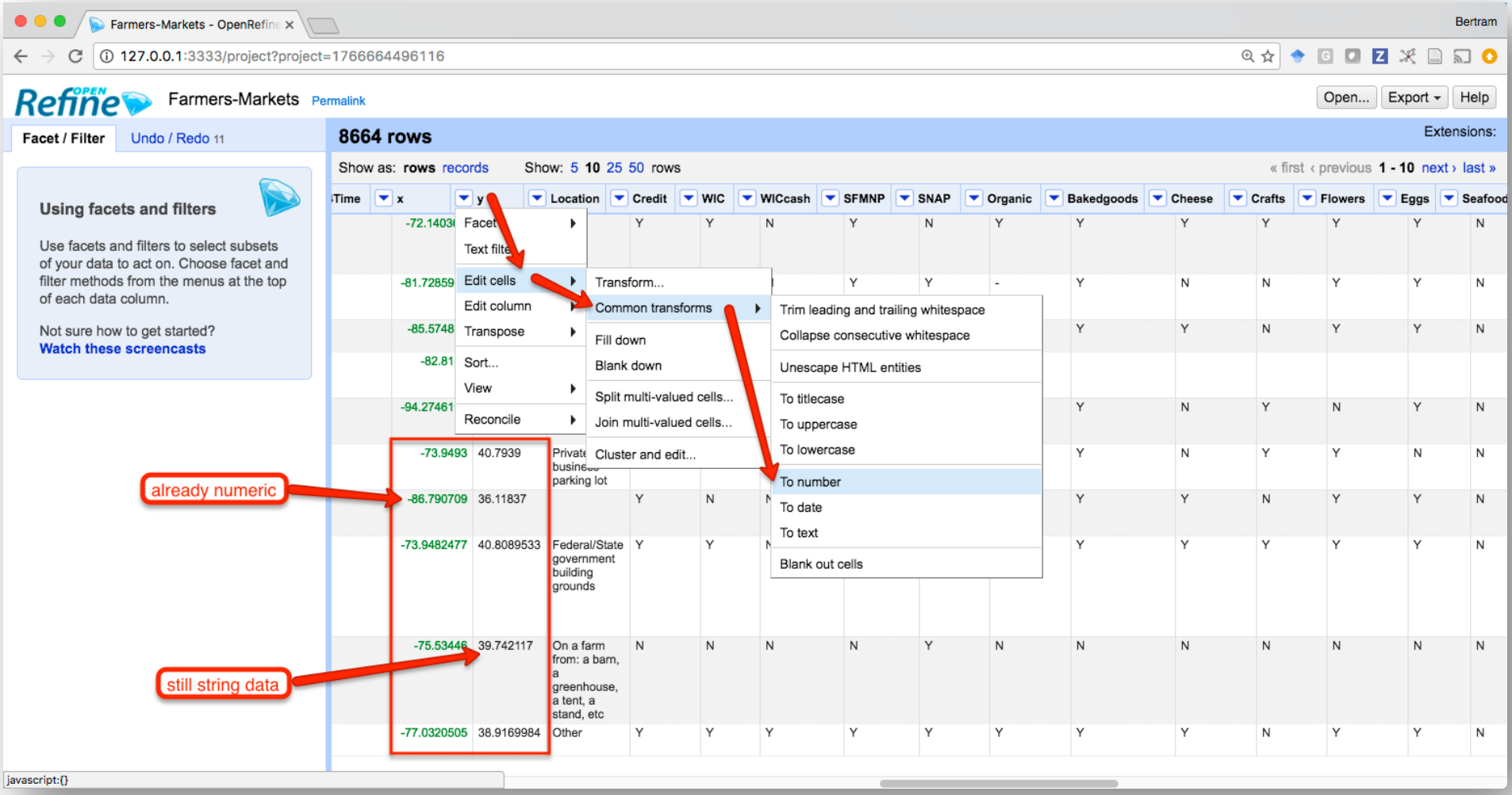
Exploring time slices: missing data …



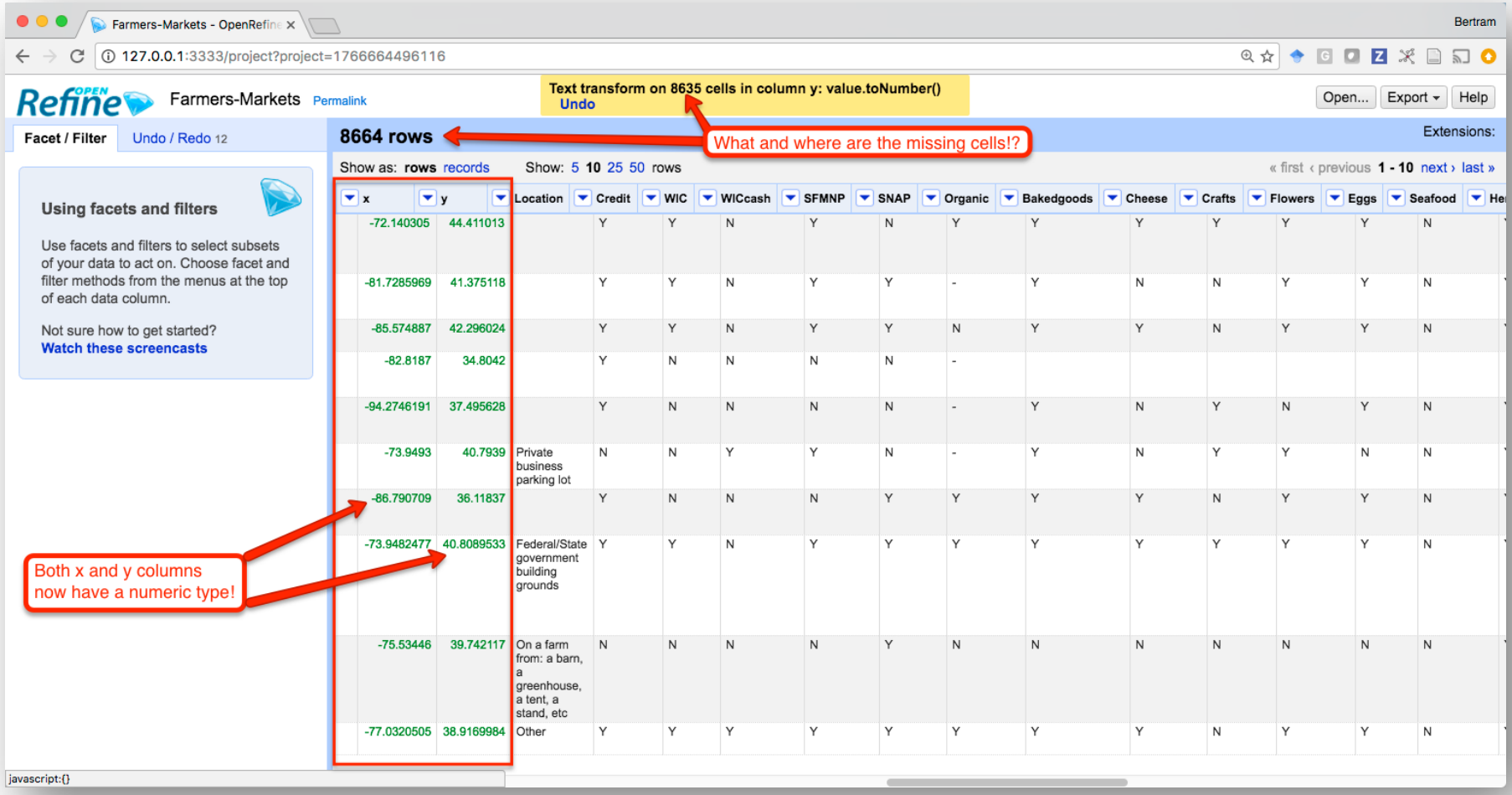
… and slices with detailed data!



Converting from **String** to **Number** …

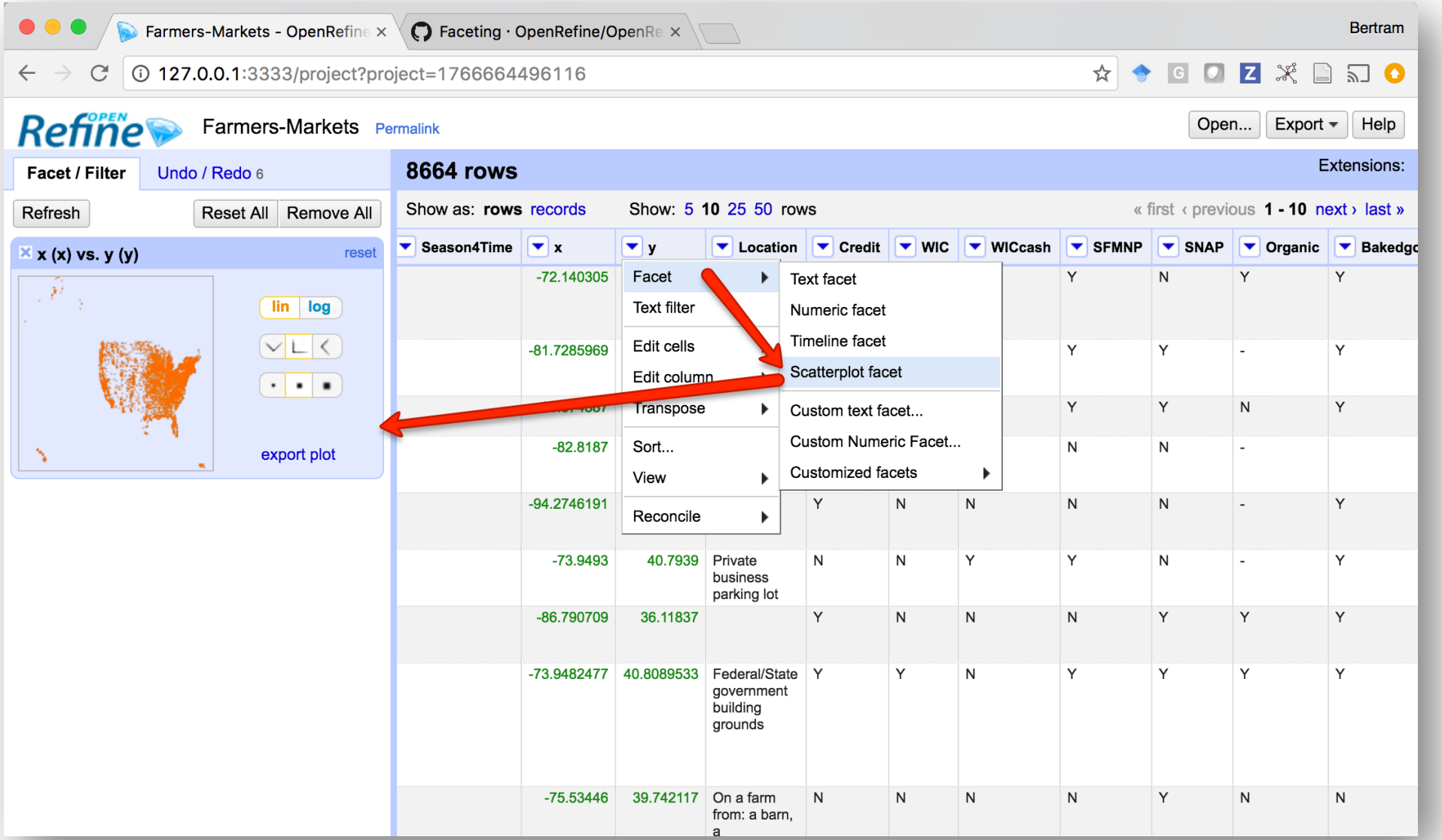


… from String to Number: Done!

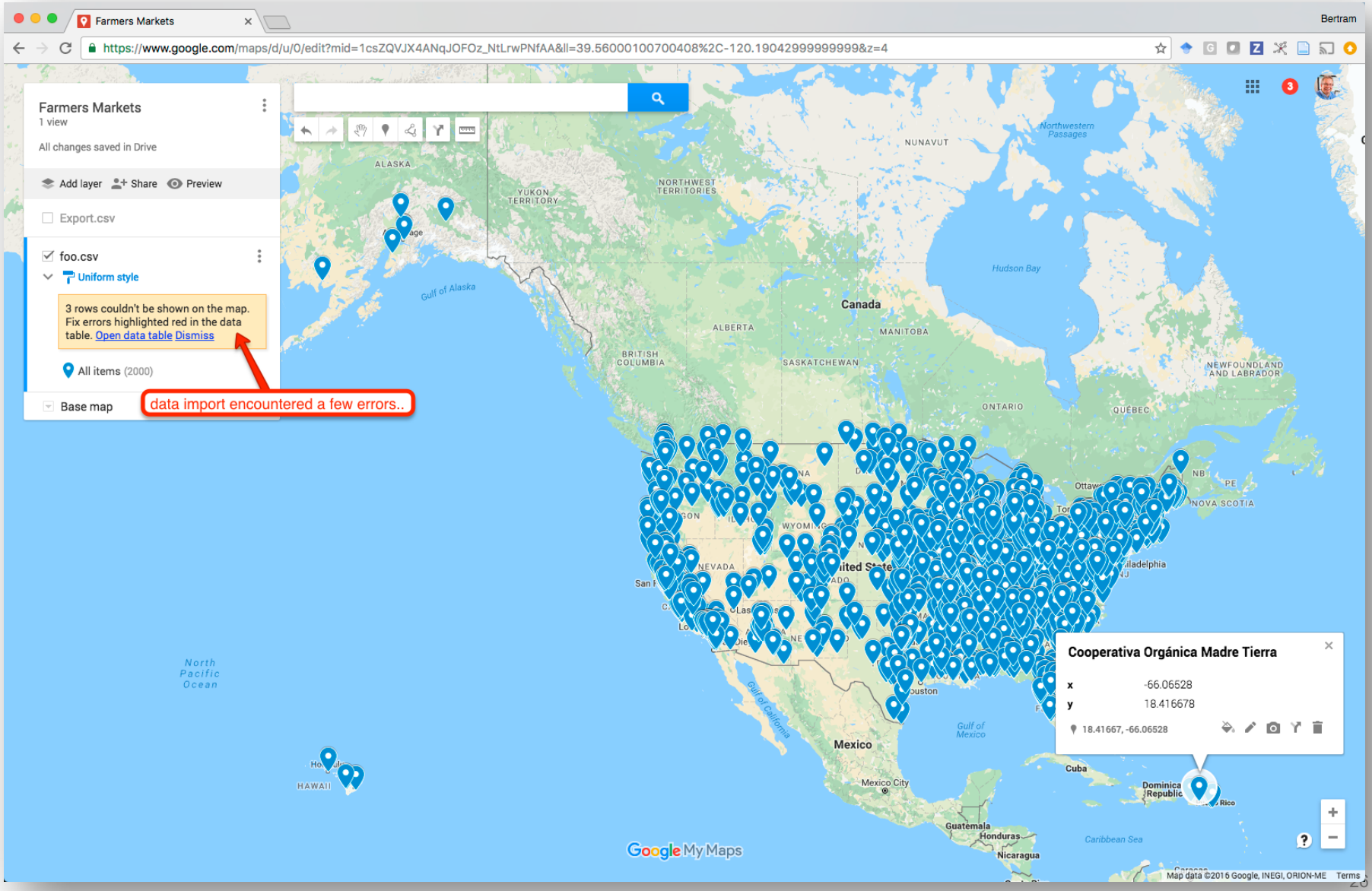


x	y	Location	Credit	WIC	WICcash	SFMNP	SNAP	Organic	Bakedgoods	Cheese	Crafts	Flowers	Eggs	Seafood
-72.140305	44.411013		Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	N
-81.7285969	41.375118		Y	Y	N	Y	Y	-	Y	N	N	Y	Y	N
-85.574887	42.296024		Y	Y	N	Y	Y	N	Y	Y	N	Y	Y	N
-82.8187	34.8042		Y	N	N	N	N	-						
-94.2746191	37.495628		Y	N	N	N	N	-	Y	N	Y	N	Y	N
-73.9493	40.7939	Private business parking lot	N	N	Y	Y	N	-	Y	N	Y	Y	N	N
-86.790709	36.11837		Y	N	N	N	Y	Y	Y	Y	N	Y	Y	N
-73.9482477	40.8089533	Federal/State government building grounds	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N
-75.53446	39.742117	On a farm from: a barn, a greenhouse, a tent, a stand, etc	N	N	N	N	Y	N	N	N	N	N	N	N
-77.0320505	38.9169984	Other	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	N

The x,y (longitude,latitude) data lets us use a "gem": **Scatterplot Facet!**



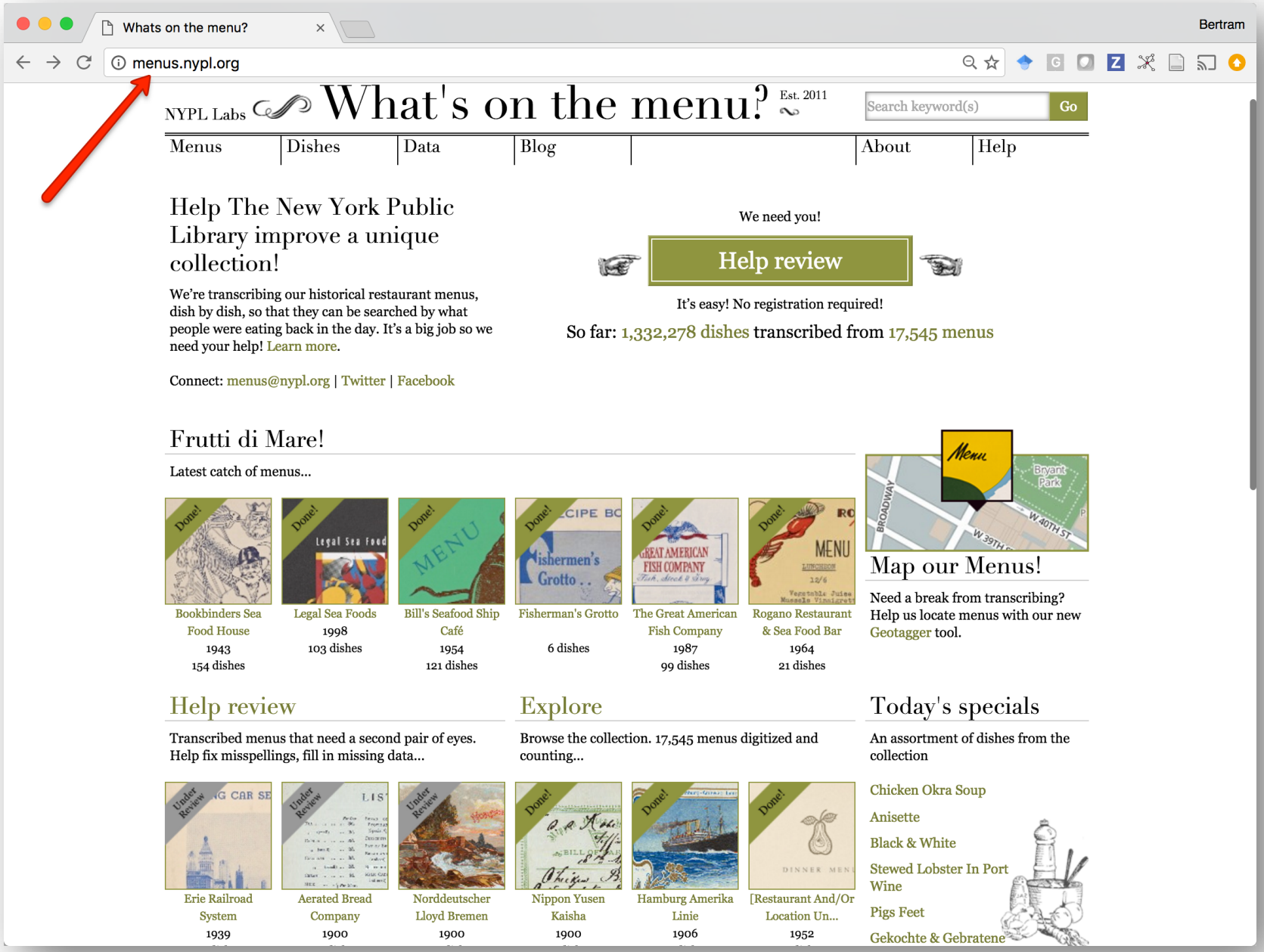
Georeferenced data & (Google) maps!

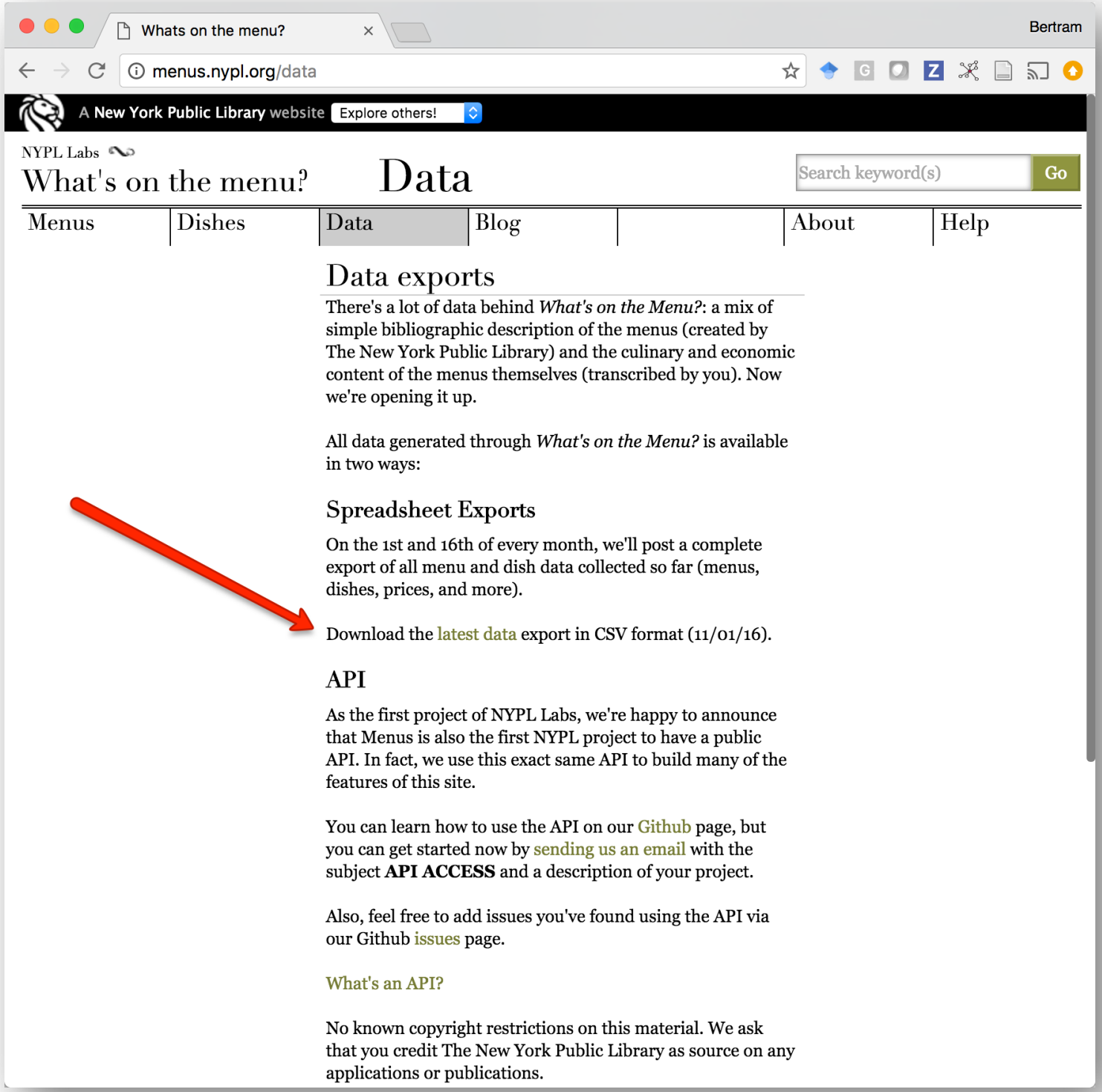


Dealing with more messy and more complex data issues …

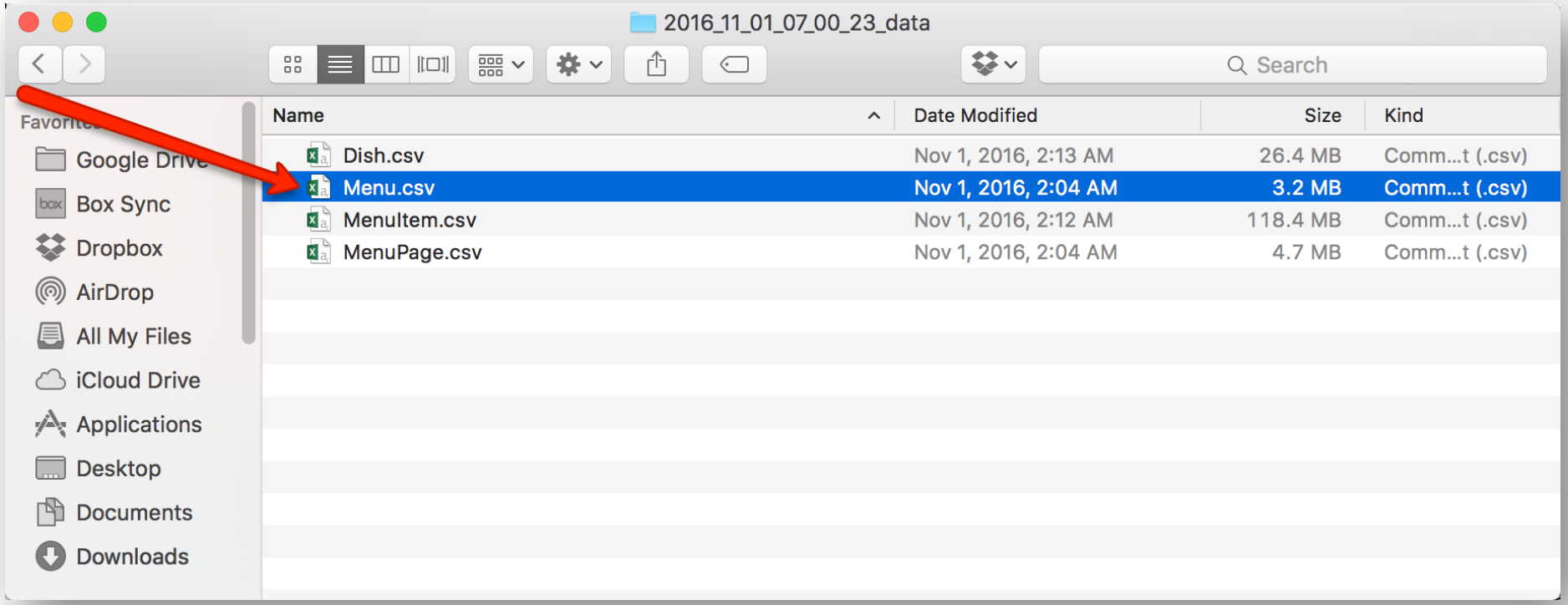
… The NYPL Menus Project!

Example: NYPL “Menu” Collection




Whats on the menu?
Bertram
menus.nypl.org/data
A New York Public Library website
Explore others!
NYPL Labs
What's on the menu?
Data
Search keyword(s)
Go
Menus
Dishes
Data
Blog
About
Help
Data exports
There's a lot of data behind What's on the Menu?: a mix of simple bibliographic description of the menus (created by The New York Public Library) and the culinary and economic content of the menus themselves (transcribed by you). Now we're opening it up.
All data generated through What's on the Menu? is available in two ways:
Spreadsheet Exports
On the 1st and 16th of every month, we'll post a complete export of all menu and dish data collected so far (menus, dishes, prices, and more).
Download the latest data export in CSV format (11/01/16).
API
As the first project of NYPL Labs, we're happy to announce that Menus is also the first NYPL project to have a public API. In fact, we use this exact same API to build many of the features of this site.
You can learn how to use the API on our Github page, but you can get started now by sending us an email with the subject API ACCESS and a description of your project.
Also, feel free to add issues you've found using the API via our Github issues page.
What's an API?
No known copyright restrictions on this material. We ask that you credit The New York Public Library as source on any applications or publications.

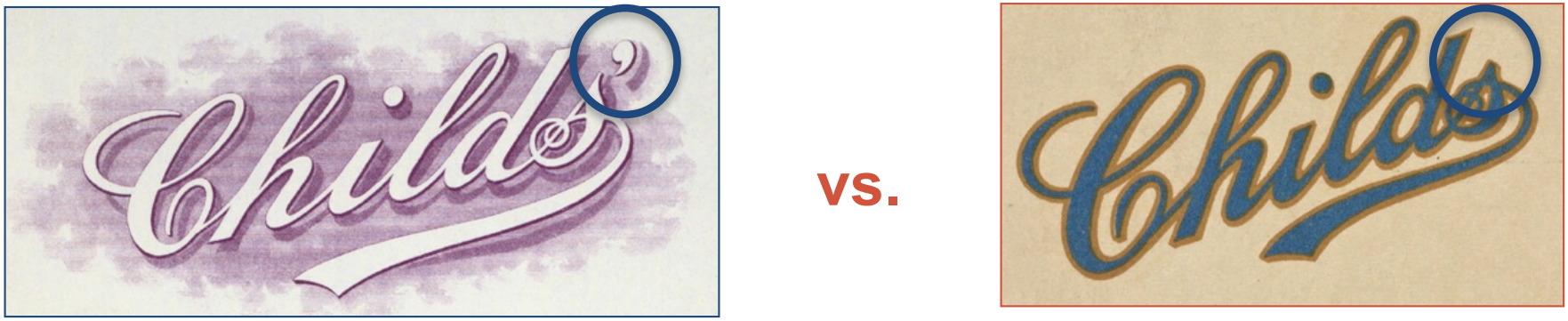
Unpacking and selection Menu.csv …



Complex Data Quality Issues: To fix or not to fix?

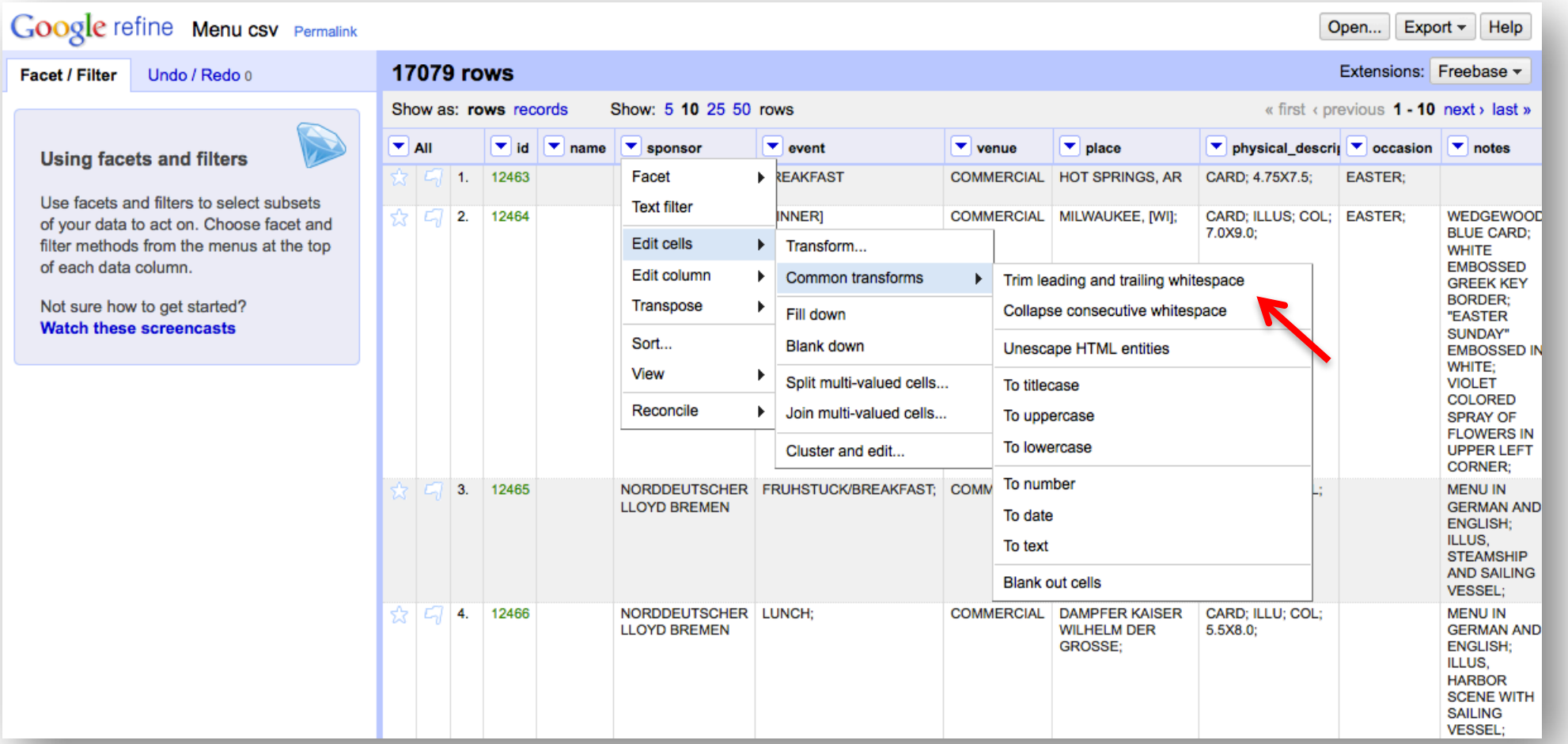
- Relying on volunteer transcription will often result in inconsistent data entry
- Even well-transcribed data is subject to challenges due to synonyms and spelling variants, etc.
- Also: entities change over time…

e.g., Childs' restaurant, originally launched by brothers Samuel and William Childs in 1889, grew to be one of the first national dining chains and dropped its apostrophe sometime after 1907.

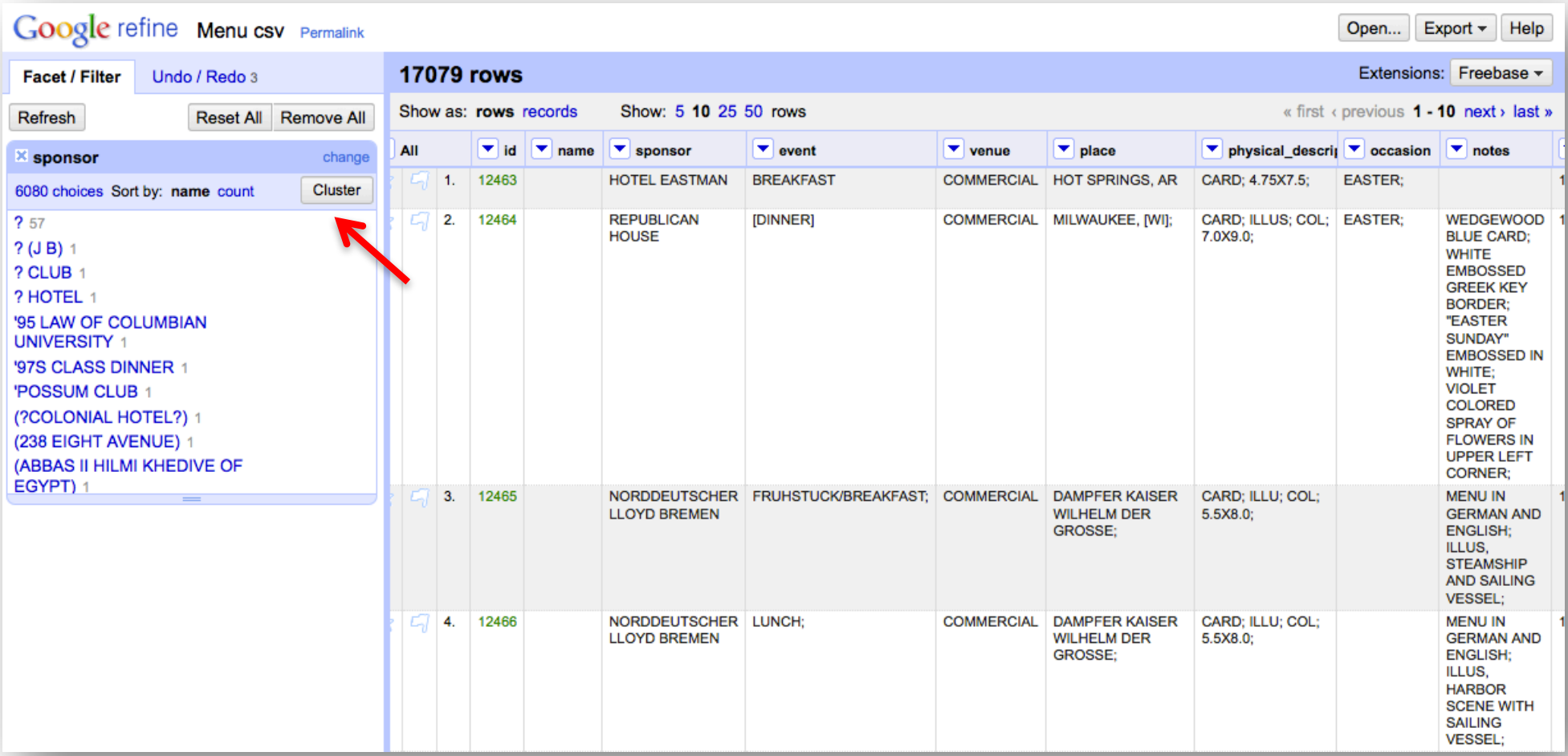


The same restaurant styled differently in 1907 (left) and 1916 (right)

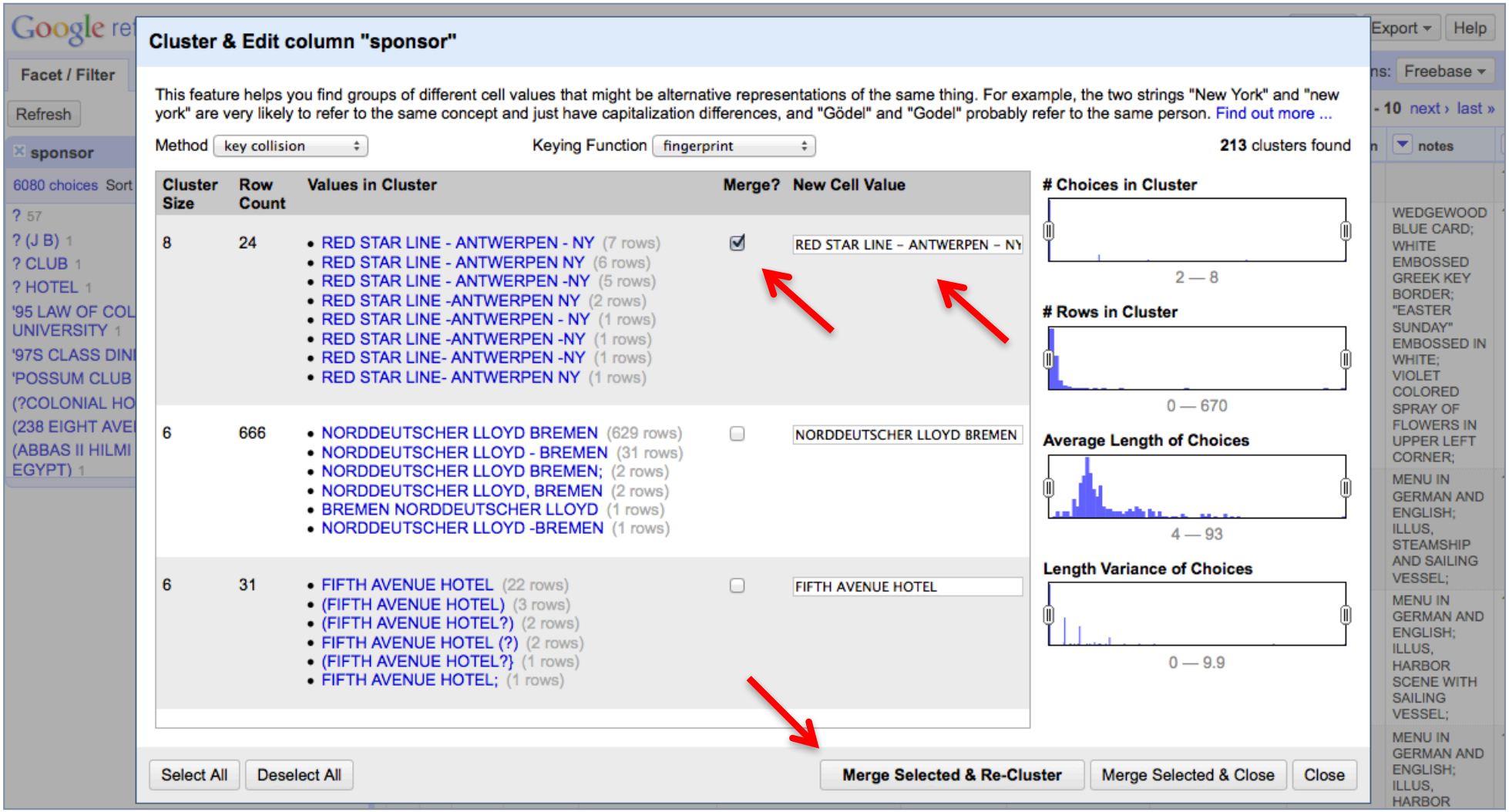
Basic Normalization



Faceting and Clustering



Faceting and Clustering



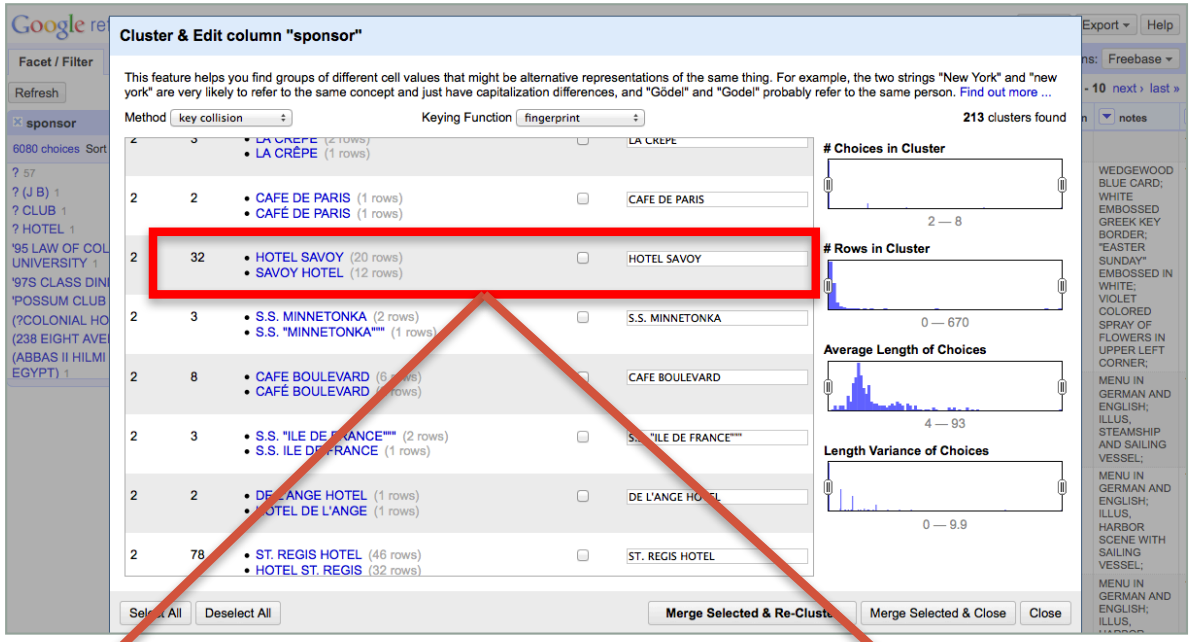
Kinds of Clustering

- Key collision (fastest, safest)
  - Fingerprint, Ngram Fingerprint = defaults
    - Match normalized strings in different ways
  - Metaphone = English pronunciation
- Nearest Neighbor
  - PPM = Partial matching
  - Levenshtein = edit distance

But beware: Clustering Caveat!

Hotel Savoy
59th St. & 5th Ave.
New York, New York

Savoy Hotel
Strand
London WC2R 0EU
United Kingdom



Summary: A First Look at OpenRefine

- Creating a New Project
- Basic Normalization
- Different Facets (text, timeline, scatterplot)
- Clustering and Mass Edits
- Operation History: Provenance

- Separate videos:
  - Installing OpenRefine
  - Advanced Operations