

Team 140 Final Report

Table of Contents

Phase I Report	2
Dataset.....	2
Target Use Cases.....	2
<i>U1</i> - Main Use Case	2
<i>U0</i> - No Data Cleaning Required	2
<i>U2</i> - Data Clearning not Sufficient.....	2
Dataset Details	2
Data Quality Problems	3
Initial Plan	7
Cleaning Steps	7
Phase II Report.....	8
Data cleaning performed.....	8
Cleaning Menu.csv file	8
Cleaning Dish.csv file.....	18
Data quality changes	22
Summary of changes.....	22
Before and After	24
ICV Queries	25
Workflow Model.....	29
Workflow W0:	29
Workflow W1: Cleaning Menu.csv file	30
Workflow W2: Cleaning Dish.csv file.....	30
Conclusions & Summary	31

Vikas Seth (Team 140)

vseth3@illinois.edu

1. Points received for initial Phase-1 submission: 96 (out of 100)
2. Our team chooses the following Phase-1 option:
 - [X] (A) No change to Phase-1 report
 - [] (B) Improved (or extended) Phase-1 report

Phase I Report

Dataset

I will be using NYPL Menu data set for the final project which is available at <http://menus.nypl.org/>. This is dataset of Restaurant Menu collection dating from 1840s collected by New York Public Library. This collection is one of the largest in the world, used by historians, chefs, novelists and every food enthusiasts. This collection mainly contains details about different Menus from the past along with details of dishes, price range, menu size, location, sponsor details of the Menus.

Target Use Cases

***U₁* - Main Use Case**

One of the interesting use case with this data-set could be about dishes. Like when a dish appeared for the first time and on which Menu, who was the sponsor and which year it appeared. For this we will need to clean this dataset to make sure that we have cleaned information about name of the dishes, Restaurants, sponsor and year information. Based on that we should be able to search this database to filter out this details for any dish.

***U₀* - No Data Cleaning Required**

To find out average number of dishes in Menu for each by each year or decade.

***U₂* - Data Cleaning not Sufficient**

Origin Language of the Menu. This information is available for some of the menus in Notes field, but it is not available for all of them. Also, there is a language column. But this information is not available there as well. So, even if we try to clean the data under notes column we will not be able to find this information for all the Menus.

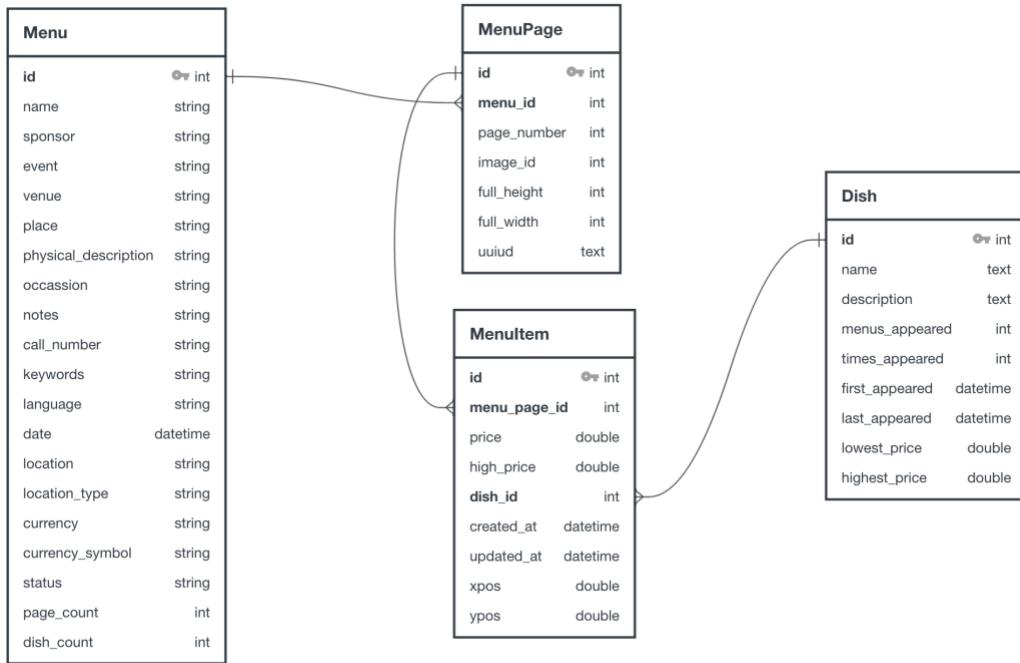
Dataset Details

This is data set for Menu details collected by NYPL with the help of Crowdsourcing. This menu collection goes back to the year 1840 and contains details of Menus like dishes, occasion, sponsor, venue type, place, location, number of dishes along with physical properties of Menu like number of pages, bind type etc. This dataset mainly contains data in following 4 files

1. **Menu** - It contains overall details of each menu like Date, Sponsor, Event, Venue, Restaurant, Number of pages, Number of dishes in the Menu

2. **Menu Page** - Unique key in this set is Page id and for each Menu along with Page Number it contains Image details and dimensions of the Menu
3. **Menu Item** - Contains details of items/dishes, price, and dish id in a Menu
4. **Dish** - It contains details of the dish including name, description, number of occurrence in Menus, Occurrence timeline and price details.

Basic database schema of this dataset is as follows



Data Quality Problems

At minimum we will need to address following data quality problems with this dataset to make it usable for use case U_1

1. **Cleaning up Dishes name:** Below is the screenshot of few examples name of the dishes we will need to clean.

Method key collision ▾ Keying Function fingerprint ▾

			• ??? (1 rows)	
			• ???? (1 rows)	
11	11		• (2) Boiled Eggs (1 rows)	<input type="checkbox"/> (2) Boiled Eggs
			• (2) Eggs Boiled (1 rows)	
			• 2 Boiled Eggs (1 rows)	
			• 2 Boiled Eggs (1 rows)	
			• 2 Boiled eggs (1 rows)	
			• 2 Eggs Boiled (1 rows)	
			• 2 Eggs, boiled (1 rows)	
			• 2 eggs boiled (1 rows)	
			• 2 eggs, boiled (1 rows)	
			• 2 eggs: boiled (1 rows)	
			• [2] BOILED EGGS (1 rows)	
10	11		• American Cheese Sandwich (2 rows)	<input type="checkbox"/>
			• AMERICAN CHEESE SANDWICH (1 rows)	
			• AMERICAN CHEESE Sandwich (1 rows)	
			• AMERICAN CHEESE sandwich (1 rows)	
			• American Cheese Sandwich (1 rows)	
			• American Cheese [Sandwich] (1 rows)	
			• American Cheese sandwich (1 rows)	
			• American cheese (sandwich) (1 rows)	
			• American cheese sandwich (1 rows)	
			• american cheese sandwich (1 rows)	
9	9		• (2) Eggs Fried (1 rows)	<input type="checkbox"/>
			• (2) Fried Eaas (1 rows)	

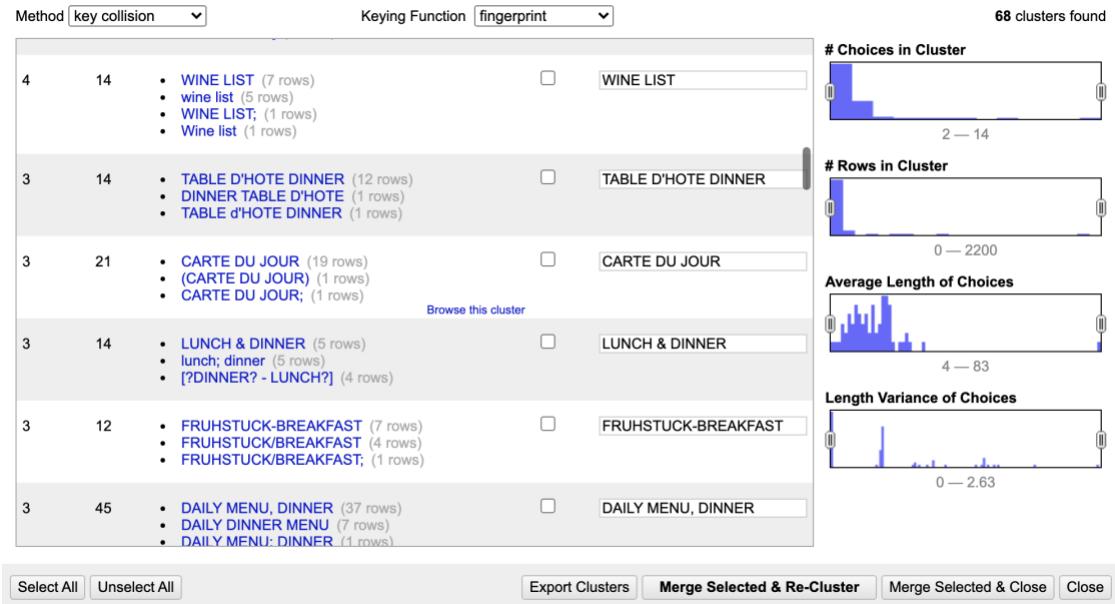
2. **Cleaning up Sponsor/Location Details:** Below is the screenshot of few examples Sponsor/Location names we need to clean up

Method key collision ▾ Keying Function ngram-fingerprint ▾ Ngram Size 2 ▾ 169 clusters found

			• Nippon Yusen Kaisha Ss Kasuga (1 rows)	
4	33		• Fifth Avenue Hotel (28 rows)	<input type="checkbox"/> Fifth Avenue Hotel
			• Fifth Avenue Hotel (?) (2 rows)	
			• Fifth Avenue Hotel? (2 rows)	
			• {Fifth Avenue Hotel?} (1 rows)	
4	791		• Waldorf Astoria (703 rows)	<input type="checkbox"/> Waldorf Astoria
			• Waldorf-Astoria (84 rows)	
			• Waldorf Astoria (2 rows)	
			• Waldorf Astoria? (2 rows)	
4	13		• Nippon Yusen Kaisha S.S.Kobe Maru (5 rows)	<input type="checkbox"/> Nippon Yusen Kaisha S.S.Kobe Maru
			• Nippon Yusen Kaisha S.S. Kobe Maru (4 rows)	
			• Nippon Yusen Kaisha S.S.Kobe Maru (2 rows)	
			• Nippon Yusen Kaisha S.S. Kobe Maru (2 rows)	
3	3		• Hotel Du Pont (1 rows)	<input type="checkbox"/> Hotel Du Pont
			• Hotel DuPont (1 rows)	
			• Hotel duPont (1 rows)	
		Browse this cluster		
3	18		• Cie Gle Transatlantique (10 rows)	<input type="checkbox"/> Cie Gle Transatlantique
			• Cie. Gle. Transatlantique (7 rows)	
			• Cie.Gle. Transatlantique (1 rows)	

Choices in Cluster
Rows in Cluster
Average Length of Choices
Length Variance of Choices

3. **Clean Event Types:** Below is the screenshots for sample event types names required cleaning



4. **Date Values:** There seems to some date values outside range will require some cleaning, like following

6599 choices Sort by: name count

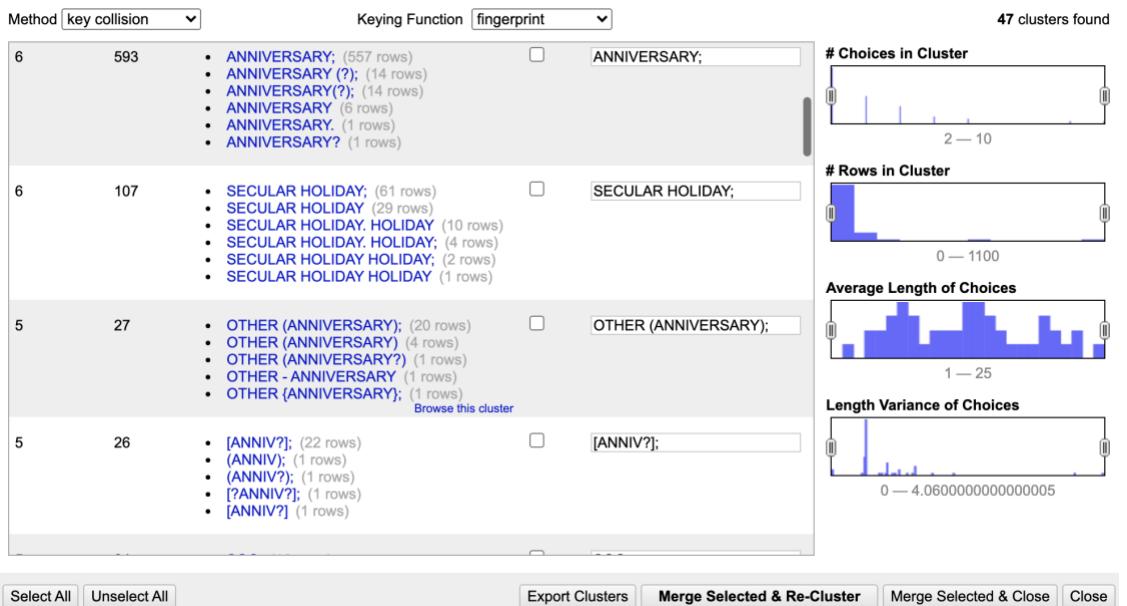
0001-01-01 2
0190-03-06 1
1091-01-27 1

5. **Additional Cleaning:** Apart from fields needed for usecase U1. There are some more fields can be cleaned and provide more details about menus. Like following

1. **Physical Details of the Menus:** Seems like following can be cleaned and we can at-least get some information about Menu build like Card, Folder, etc.

physical_description		change
6147 choices Sort by: name count		Cluster
34.5x25cm 40		
28x20.5cm folded; 28x41cm open 46		
34x23cm 45		
34x24cm folded; 34x48cm open 44		
CARD; ILLUS; COL; 5X8; 43		
24x16.5cm folded; 24x33cm open 41		
CARD; COL; ILLUS; 5.5 X 8; 40		
19.5x14cm 39		
BROADSHEET; COL; ILLUS; 10.25		
X 9.5; 39		
CARD; ILLUS; 5 X 8; 39		
CARD;ILL;COL;5X8; 39		
34x23cm 39		

2. *Occasion*: Occasion details can also be cleaned to get the details if Menu was used for certain Occasion



3. *Additional Cleaning for Dishes*: There are certain dishes having 0 count for Menu occurrence and not year information. We should be able to drop those from our analysis

name	description	menus_appear	times_appeared	first_appeared	last_appeared	lowest_pr
0.75		1	0	0	0	0
80		10	10	0	0	0
* au gratin		1	0	1900	1900	0.4
* with mushrooms		1	0	1900	1900	1.6
(Meat Balls with Chicken, Mush-		1	0	0	0	0
(per bowl)		1	0	0	0	0
Almond		1	0	0	0	0
Chicken Chow Mein		1	0	0	0	1.1
Chicken Chow Mein (For 2)		1	0	0	0	0
Gaw Mein		1	0	0	0	0.45
Gravy		1	0	0	0	0
Mushroom		1	0	0	0	0
nuts and Vegetable.)		1	0	0	0	0
roon, Bambooshoots, Waterchest-		1	0	0	0	0
Sandwich: Sliced Turkey, on home-made European type dark bread with lettuce and tomatoes, organically grown at Trapp Gardens		1	0	1966	1966	1.25
Sandwich: Sliced Turkey, on home-made European type dark bread, served open with lettuce and tomatoes, organically grown at Trapp Gardens		1	0	0	0	0
Sliced Turkey Sandwich, on home-made European type dark bread, served open with lettuce and tomatoes, organically grown at Trapp Gardens		1	0	0	0	0
A Real Treat-PIZZABURGER with Potato Chips		1	0	1969	1969	0.7
All White Meat Sliced Chicken Cold Cut Platter		1	0	1959	1959	1.75
Breakfast No. 6 - Fruit, Fruit Juice or Cereal, Choice of French Toast with Syrup or Jelly or Wheat or Corn Cakes with Honey or Syrup, Coffee, Tea, Milk		1	0	1945	1945	0
Codfish balls		1	0	0	0	0
Darne de Saumon grillé Bearnaise		1	0	1954	1954	0
Dry Toast		1	0	0	0	0
fresh salad		1	0	1987	1987	0
frisch gepressster Orangensaft		1	0	1988	1988	0
Green Peas		1	0	0	0	0

Initial Plan

Cleaning Steps

1. Convert data to appropriate formats like Date, number or text.
2. Using OpenRefine clean the text fields for leading and trailing spaces and also clean for consecutive spaces
3. Remove any special characters using OpenRefine and Regular Expression
4. Using OpenRefine use clustering methods to clean the fields like Dishes name, Sponsor, Location, Event, Occasion
5. Trying cleaning Physical Description of the Menu using Regular Expression to find Build Type of the Menu
6. Using SQLite build the database schema for the dataset to check for ICs like:
 1. All ids should be unique for Dishes, Menus, Menu Pages
 2. ids should not be null for Dishes, Menus, Menu Pages
 3. Check for the dishes where last appeared or first appeared is 0
 4. Lowest Price should not be greater than Highest Price
 5. First Appeared should not be greater than last appeared
 6. Page count in a Menu should not be null or 0

Phase II Report

Data cleaning performed

To meet the use case U_1 I decided to clean the following columns in the respective files using OpenRefine.

Cleaning Menu.csv file

Column: sponsor

As in the U_1 we are planning to identify who was the sponsor of a dish at a particular time. Hence cleaning of this column was required to get correct sponsor details.

Step 1: Trim Leading and Trailing White Spaces

Step 2: Collapse consecutive white spaces

Step 3: GREL expression to remove special characters

The screenshot shows the 'Custom text transform' dialog in OpenRefine. The 'Language' dropdown is set to 'General Refine Expression Language (GREL)'. The 'Expression' field contains the GREL code: `value.replace(/[\t\n\r\f\v]/g, '')`. A green 'G' icon indicates no syntax errors. The 'Preview' tab is selected, showing the transformation of a list of sponsor names. The original values are on the left, and the transformed values are on the right. The transformed values have all special characters removed. At the bottom, there are options for 'On error': 'keep original' (radio button selected), 'set to blank', and 'store error'. There is also a checkbox for 'Re-transform up to [10] times until no change'. At the bottom left are 'OK' and 'Cancel' buttons.

row	value	value.replace(/[\t\n\r\f\v]/g, '')
1.	HOTEL EASTMAN	HOTEL EASTMAN
2.	REPUBLICAN HOUSE	REPUBLICAN HOUSE
3.	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD BREMEN
4.	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD BREMEN
5.	NORDDEUTSCHER LLOYD BREMEN	NORDDEUTSCHER LLOYD BREMEN
6.	CANADIAN PACIFIC RAILWAY COMPANY	CANADIAN PACIFIC RAILWAY COMPANY
7.	HOTEL NETHERLAND	HOTEL NETHERLAND

Step 4: Cluster and Edit Column - fingerprint

Cluster & Edit column "sponsor"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	key collision	Keying Function	fingerprint	307 clusters found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	24	<ul style="list-style-type: none"> RED STAR LINE - ANTWERPEN - NY (7 rows) RED STAR LINE - ANTWERPEN NY (6 rows) RED STAR LINE - ANTWERPEN -NY (5 rows) RED STAR LINE -ANTWERPEN -NY (2 rows) RED STAR LINE -ANTWERPEN - NY (1 rows) RED STAR LINE -ANTWERPEN -NY (1 rows) RED STAR LINE-ANTWERPEN -NY (1 rows) RED STAR LINE-ANTWERPEN NY (1 rows) 	<input type="checkbox"/>	RED STAR LINE - ANTWERPE
6	667	<ul style="list-style-type: none"> NORDDEUTSCHER LLOYD BREMEN (507 rows) Norddeutscher Lloyd Bremen (125 rows) NORDDEUTSCHER LLOYD - BREMEN (31 rows) NORDDEUTSCHER LLOYD, BREMEN (2 rows) BREMEN NORDDEUTSCHER LLOYD (1 rows) NORDDEUTSCHER LLOYD -BREMEN (1 rows) 	<input type="checkbox"/>	NORDDEUTSCHER LLOYD BI
5	73	<ul style="list-style-type: none"> Delmonicos (50 rows) Delmonico's (10 rows) DELMONICO'S (9 rows) DELMONICOS (3 rows) Delmonicos. (1 rows) 	<input type="checkbox"/>	Delmonicos
5	18	<ul style="list-style-type: none"> Hotel Imperial (5 rows) IMPERIAL HOTEL (5 rows) 	<input type="checkbox"/>	Hotel Imperial

Choices in Cluster
Histogram showing the distribution of the number of choices per cluster. Range: 2 — 8.

Rows in Cluster
Histogram showing the distribution of the number of rows per cluster. Range: 0 — 700.

Average Length of Choices
Histogram showing the distribution of the average length of choices per cluster. Range: 3 — 93.

Length Variance of Choices
Histogram showing the distribution of the length variance of choices per cluster. Range: 0 — 8.67.

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Step 5: Cluster and Edit Column - ngram-fingerprint

Cluster & Edit column "sponsor"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	key collision	Keying Function	ngram-fingerprint	Ngram Size	2	75 clusters found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value		
3	4	<ul style="list-style-type: none"> RED STAR LINE - S.S.FRIESLAND (2 rows) RED STAR LINE SS FRIESLAND (1 rows) RED STAR LINE-SS FRIESLAND (1 rows) 	<input type="checkbox"/>	RED STAR LINE - S.S.FRIESL		
3	4	<ul style="list-style-type: none"> U.S.S. RALEIGH (2 rows) U.S.S.RALEIGH (1 rows) U.S.S.S.RALEIGH (1 rows) 	<input type="checkbox"/>	U.S.S. RALEIGH		
3	8	<ul style="list-style-type: none"> Hofbrau Haus (5 rows) Hofbrauhau (2 rows) Hof Brau Haus (1 rows) 	<input type="checkbox"/>	Hofbrau Haus		
3	10	<ul style="list-style-type: none"> NIPPON YUSEN KAISHA - S.S.KOBE MARU (5 rows) NIPPON YUSEN KAISHA - S.S. KOBE MARU (4 rows) NIPPON YUSEN KAISHA - S.S. KOBE MARU (1 rows) 	<input type="checkbox"/>	NIPPON YUSEN KAISHA - S.S.		
2	58	<ul style="list-style-type: none"> ? (57 rows) L (1 rows) 	<input type="checkbox"/>	?		

Choices in Cluster
Histogram showing the distribution of the number of choices per cluster. Range: 2 — 3.

Rows in Cluster
Histogram showing the distribution of the number of rows per cluster. Range: 0 — 790.

Average Length of Choices
Histogram showing the distribution of the average length of choices per cluster. Range: 1 — 49.

Length Variance of Choices
Histogram showing the distribution of the length variance of choices per cluster. Range: 0 — 2.

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Step 6: Redo of Trim Leading and Trailing White Spaces & Collapse consecutive white spaces

Column: event

As in the **U₁** we are planning to identify which was the event of a dish at a particular time. Hence cleaning of this column was required to get correct event details.

Step 1: Trim Leading and Trailing White Spaces

Step 2: Collapse consecutive white spaces

Step 3: GREL expression to remove special characters

Custom text transform on column event

Expression: value.replace(/[\r\n\t\f]/g, " ")

Language: General Refine Expression Language (GREL)

No syntax error.

row	value	value.replace(/[\r\n\t\f]/g, " ")
1.	BREAKFAST	BREAKFAST
2.	[DINNER]	DINNER
3.	FRUHSTUCK/BREAKFAST;	FRUHSTUCK/BREAKFAST
4.	LUNCH;	LUNCH
5.	DINNER;	DINNER
6.	[DINNER]	DINNER
7.	CUSPED	CUSPED

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 4: GREL expression to replace '?' at the end of '- TENTATIVE'

Custom text transform on column event

Expression: value.replace(/\?\$/g, " - TENTATIVE")

Language: General Refine Expression Language (GREL)

No syntax error.

row	value	value.replace(/\?\$/g, " - TENTATIVE")
1.	BREAKFAST	BREAKFAST
2.	DINNER	DINNER
3.	FRUHSTUCK/BREAKFAST	FRUHSTUCK/BREAKFAST
4.	LUNCH	LUNCH
5.	DINNER	DINNER
6.	DINNER	DINNER
7.	CUSPED	CUSPED

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 5: Removing additional '?'

Custom text transform on column event

Expression: value.replace(/\?/, "")

Language: General Refine Expression Language (GREL)

No syntax error.

Preview	History	Starred	Help
row value	value.replace(/\?/, "")		
1. BREAKFAST	BREAKFAST		
2. DINNER	DINNER		
3. FRUHSTUCK/BREAKFAST	FRUHSTUCK/BREAKFAST		
4. LUNCH	LUNCH		
5. DINNER	DINNER		
6. DINNER	DINNER		
7. DINNER	DINNER		

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 6: Cluster and Edit Column – fingerprint

Cluster & Edit column "event"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: key collision Keying Function: fingerprint 56 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	626	<ul style="list-style-type: none"> LUNCH (537 rows) lunch (55 rows) Lunch (33 rows) LUNCH. (1 rows) 	<input type="checkbox"/>	LUNCH
4	2134	<ul style="list-style-type: none"> DINNER (1921 rows) dinner (142 rows) Dinner (70 rows) DINNER, (1 rows) 	<input type="checkbox"/>	DINNER
3	14	<ul style="list-style-type: none"> TABLE D'HOTE DINNER (12 rows) DINNER TABLE D'HOTE (1 rows) TABLE d'HOTE DINNER (1 rows) 	<input type="checkbox"/>	TABLE D'HOTE DINNER
3	5	<ul style="list-style-type: none"> Afternoon tea (2 rows) afternoon tea (2 rows) AFTERNOON TEA (1 rows) 	<input type="checkbox"/>	Afternoon tea
3	14	<ul style="list-style-type: none"> WINE LIST (8 rows) wine list (5 rows) Wine list (1 rows) 	<input type="checkbox"/>	WINE LIST
0	444	SUPPER (405 rows)	<input type="checkbox"/>	SUPPER

Choices in Cluster: 2 — 4

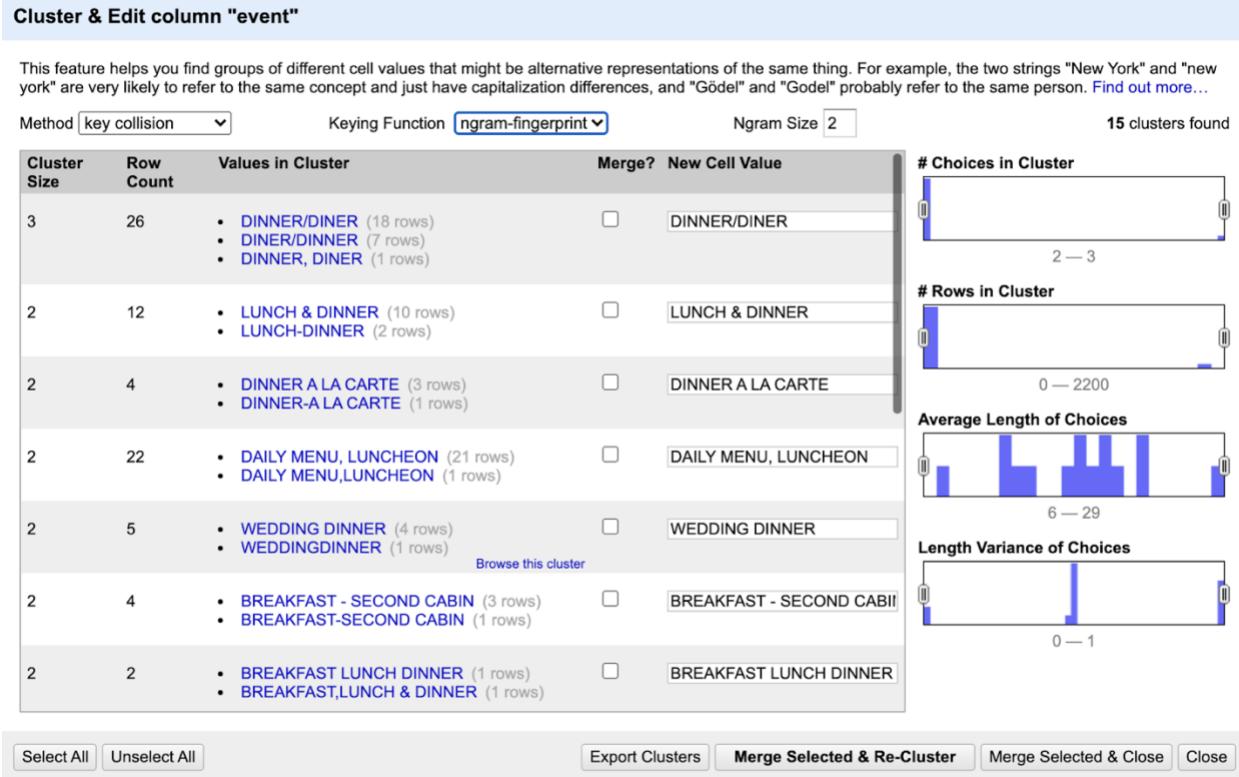
Rows in Cluster: 0 — 2200

Average Length of Choices: 4 — 81

Length Variance of Choices: 0 — 1

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Step 7: Cluster and Edit Column – ngram-fingerprint



Step 8: Redo of Trim Leading and Trailing White Spaces & Collapse consecutive white spaces

Column: occasion

As in the U_1 we are planning to identify which was the occasion of a dish at a particular time. Hence cleaning of this column was required to get correct occasion details.

Step 1: Trim Leading and Trailing White Spaces

Step 2: Collapse consecutive white spaces

Step 3: GREL expression to remove special characters

Custom text transform on column occasion

Expression: value.replace(/;|\n|\\|\(|\)|\||\"|\|\|/, "")

Language: General Refine Expression Language (GREL)

No syntax error.

Preview	History	Starred	Help
42. DAILY	DAILY		
43. DAILY	DAILY		
44. DAILY	DAILY		
45. DAILY	DAILY		
46. DAILY	DAILY		
47. DAILY	DAILY		
48. ANNIVERSARY;	ANNIVERSARY		

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 4: GREL expression to replace '?' at the end of ' - MAYBE'

Custom text transform on column occasion

Expression: value.replace(/ \?\$/|\?\$/ , " - MAYBE")

Language: General Refine Expression Language (GREL)

No syntax error.

Preview	History	Starred	Help
42. DAILY	DAILY		
43. DAILY	DAILY		
44. DAILY	DAILY		
45. DAILY	DAILY		
46. DAILY	DAILY		
47. DAILY	DAILY		
48. ANNIVERSARY	ANNIVERSARY		

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 5: Removing additional '?'

Custom text transform on column occasion

Expression: value.replace(/\?/, "")

Language: General Refine Expression Language (GREL)

No syntax error.

Preview	History	Starred	Help
42. DAILY	DAILY		
43. DAILY	DAILY		
44. DAILY	DAILY		
45. DAILY	DAILY		
46. DAILY	DAILY		
47. DAILY	DAILY		
48. ANNIVERSARY	ANNIVERSARY		
49. null			

Error: replace expects 3 strings, or 1 string, 1 regex, and 1 string

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 6: Cluster and Edit Column – fingerprint

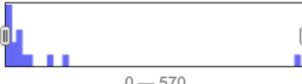
Cluster & Edit column "occasion"

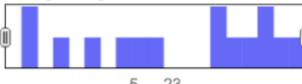
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	key collision	Keying Function	fingerprint	15 clusters found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	4	<ul style="list-style-type: none"> • OTHER ANNUAL EVENT (2 rows) • OTHER {ANNUAL EVENT} (1 rows) • OTHER,ANNUAL EVENT (1 rows) 	<input type="checkbox"/>	OTHER ANNUAL EVENT
3	75	<ul style="list-style-type: none"> • PATRIOTIC HOLIDAY (73 rows) • PATRIOTIC HOLIDAY. (1 rows) • PATRIOTIC HOLIDAY.HOLIDAY (1 rows) 	<input type="checkbox"/>	PATRIOTIC HOLIDAY
3	26	<ul style="list-style-type: none"> • OTHER ANNIVERSARY (24 rows) • OTHER - ANNIVERSARY (1 rows) • OTHER {ANNIVERSARY} (1 rows) 	<input type="checkbox"/>	OTHER ANNIVERSARY
3	107	<ul style="list-style-type: none"> • SECULAR HOLIDAY (90 rows) • SECULAR HOLIDAY.HOLIDAY (14 rows) • SECULAR HOLIDAY HOLIDAY (3 rows) 	<input type="checkbox"/>	SECULAR HOLIDAY
2	24	<ul style="list-style-type: none"> • OTHER SOC - MAYBE (23 rows) • OTHER, SOC - MAYBE (1 rows) 	<input type="checkbox"/>	OTHER SOC - MAYBE
2	3	<ul style="list-style-type: none"> • SECULAR HOLIDAY.HOLIDAY (2 rows) • SECULAR HOLIDAYHOLIDAY (1 rows) 	<input type="checkbox"/>	SECULAR HOLIDAY.HOLIDAY
2	2	• OTHER DAILY DATED MENU (1 rows)	<input type="checkbox"/>	OTHER DAILY DATED MENU

Choices in Cluster

 2 — 3

Rows in Cluster

 0 — 570

Average Length of Choices

 5 — 23

Length Variance of Choices

 0 — 4.03

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Step 7: Cluster and Edit Column – ngram-fingerprint

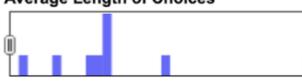
Cluster & Edit column "occasion"

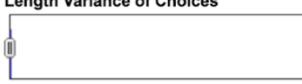
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	key collision	Keying Function	ngram-fingerprint	Ngram Size	2	9 clusters found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value		
2	2	<ul style="list-style-type: none"> • COMPL FOR MAHARAJA SCINDIYA (1 rows) • COMPLFOR MAHARAJA SCINDIYA (1 rows) 	<input type="checkbox"/>	COMPL FOR MAHARAJA SCINDIYA		
2	6	<ul style="list-style-type: none"> • OTHER ANNUAL EVENT (4 rows) • OTHER,ANNUAL EVENT (2 rows) 	<input type="checkbox"/>	OTHER ANNUAL EVENT		
2	243	<ul style="list-style-type: none"> • DAILY MENU (242 rows) • DAILYMENU (1 rows) 	<input type="checkbox"/>	DAILY MENU		
2	2	<ul style="list-style-type: none"> • OTHER COMMEMORATION (1 rows) • OTHER,Commemoration (1 rows) 	<input type="checkbox"/>	OTHER COMMEMORATION		
2	2	<ul style="list-style-type: none"> • OTHER ANNUAL BANQUET (1 rows) • OTHER,Annual Banquet (1 rows) 	<input type="checkbox"/>	OTHER ANNUAL BANQUET		
2	3	<ul style="list-style-type: none"> • OTHER ANNUAL MEETING (2 rows) • OTHER,Annual Meeting (1 rows) 	<input type="checkbox"/>	OTHER ANNUAL MEETING		
2	10	<ul style="list-style-type: none"> • RELIG. HOLIDAY (9 rows) • RELIG.HOLIDAY (1 rows) 	<input type="checkbox"/>	RELIG. HOLIDAY		

Rows in Cluster

 0 — 250

Average Length of Choices

 9 — 44

Length Variance of Choices

 0 — 0.5

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Step 8: Redo of Trim Leading and Trailing White Spaces & Collapse consecutive white spaces

Column: date

As in the U_1 we are planning to identify date of first or last occurrence of a dish. Hence cleaning of this column was required to get correct details.

Step 1: Trim Leading and Trailing White Spaces

Step 2: Collapse consecutive white spaces

Step 3: Converting into ISO date format

Custom text transform on column date

Expression: `toString(toDate(value), "yyyy-MM-dd")`

Language: General Refine Expression Language (GREL)

No syntax error.

row	value	transformed value
1.	1900-04-15	1900-04-15
2.	1900-04-15	1900-04-15
3.	1900-04-16	1900-04-16
4.	1900-04-16	1900-04-16
5.	1900-04-16	1900-04-16
6.	1900-04-16	1900-04-16

On error:

- keep original
- set to blank
- store error

Re-transform up to times until no change

OK Cancel

Column: location

As in the U_1 we are planning to identify which was the location of a dish at a particular time. Hence cleaning of this column was required to get correct location details.

Step 1: Trim Leading and Trailing White Spaces

Step 2: Collapse consecutive white spaces

Step 3: GREL expression to remove special characters

Custom text transform on column location

Expression: value.replace(/;|\{\|\}|\\|!|"/|\//, "")

Language: General Refine Expression Language (GREL)

No syntax error.

row	value	value.replace(/; \{\ \} \\ ! "/ \//, "")
1.	Hotel Eastman	Hotel Eastman
2.	Republican House	Republican House
3.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
4.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
5.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
6.	Canadian Pacific Railway Company	Canadian Pacific Railway Company

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 4: GREL expression to replace ‘?’ at the end of ‘ - TENTATIVE’

Custom text transform on column location

Expression: value.replace(/ \?\$/\?\$/ , " - TENTATIVE")

Language: General Refine Expression Language (GREL)

No syntax error.

row	value	value.replace(/ \?\$/\?\$/ , " - TENTATIVE")
1.	Hotel Eastman	Hotel Eastman
2.	Republican House	Republican House
3.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
4.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
5.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
6.	Canadian Pacific Railway Company	Canadian Pacific Railway Company

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 5: Removing additional ‘?’

Custom text transform on column location

Expression: value.replace(/\?/, "")

Language: General Refine Expression Language (GREL)

No syntax error.

row	value	value.replace(/\?/, "")
1.	Hotel Eastman	Hotel Eastman
2.	Republican House	Republican House
3.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
4.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
5.	Norddeutscher Lloyd Bremen	Norddeutscher Lloyd Bremen
6.	Canadian Pacific Railway Company	Canadian Pacific Railway Company

On error: keep original set to blank store error

Re-transform up to 10 times until no change

OK Cancel

Step 6: Cluster and Edit Column – fingerprint

Cluster & Edit column "location"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	key collision	Keying Function	fingerprint	174 clusters found
3	5	<ul style="list-style-type: none"> The Merchant's Club (3 rows) The Merchants Club (1 rows) The Merchants' Club (1 rows) 	<input type="checkbox"/> The Merchant's Club	# Choices in Cluster 2 — 3
3	5	<ul style="list-style-type: none"> Merchant's Club (2 rows) Merchants' Club (2 rows) Merchants Club (1 rows) 	<input type="checkbox"/> Merchant's Club	# Rows in Cluster 0 — 790
3	13	<ul style="list-style-type: none"> St. Denis Hotel (11 rows) Hotel St. Denis (1 rows) St. Denis Hotel (1 rows) 	<input type="checkbox"/> St. Denis Hotel	Average Length of Choices 5 — 93
3	19	<ul style="list-style-type: none"> Hotel Imperial (9 rows) Imperial Hotel (9 rows) Impérial Hotel (1 rows) 	<input type="checkbox"/> Hotel Imperial	Length Variance of Choices 0 — 9.43
3	6	<ul style="list-style-type: none"> American Medical Editor's Association (3 rows) American Medical Editors' Association (2 rows) American Medical Editors Association (1 rows) 	<input type="checkbox"/> American Medical Editor's Assoc	
2	14	<ul style="list-style-type: none"> Hotel Metropole (9 rows) Metropole Hotel (5 rows) 	<input type="checkbox"/> Hotel Metropole	
2	3	<ul style="list-style-type: none"> The Merchant's Club Of Chicago (2 rows) The Merchants Club Of Chicago (1 rows) 	<input type="checkbox"/> The Merchant's Club Of Chicag	

[Select All](#) [Unselect All](#) [Export Clusters](#) [Merge Selected & Re-Cluster](#) [Merge Selected & Close](#) [Close](#)

Step 7: Cluster and Edit Column – ngram-fingerprint

Cluster & Edit column "location"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	key collision	Keying Function	ngram-fingerprint	Ngram Size 2	67 clusters found
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Choices in Cluster
3	4	<ul style="list-style-type: none"> U.S.S. Raleigh (2 rows) U.S.S.Raleigh (1 rows) U.S.S.S.Raleigh (1 rows) 	<input type="checkbox"/>	U.S.S. Raleigh	2 — 3
3	8	<ul style="list-style-type: none"> Hofbrau Haus (5 rows) Hofbrauhaus (2 rows) Hof Brau Haus (1 rows) 	<input type="checkbox"/>	Hofbrau Haus	# Rows in Cluster 0 — 790
3	13	<ul style="list-style-type: none"> Nippon Yusen Kaisha S.S.Kobe Maru (7 rows) Nippon Yusen Kaisha S.S. Kobe Maru (4 rows) Nippon Yusen Kaisha S.S. Kobe Maru (2 rows) 	<input type="checkbox"/>	Nippon Yusen Kaisha S.S.Kobe	Average Length of Choices 4 — 47
2	2	<ul style="list-style-type: none"> Societa La Piemontese (1 rows) Societa'l'a Piemontese (1 rows) 	<input type="checkbox"/>	Societa La Piemontese	Length Variance of Choices 0 — 2
2	3	<ul style="list-style-type: none"> Hotel DuPont (2 rows) Hotel Du Pont (1 rows) 	<input type="checkbox"/>	Hotel DuPont	
2	6	<ul style="list-style-type: none"> A.H. Meyer Rathskeller (3 rows) A.H.Meyer Rathskeller (3 rows) 	<input type="checkbox"/>	A.H. Meyer Rathskeller	
2	15	<ul style="list-style-type: none"> Nippon Yusen Kaisha S.S.Kosuge (10 rows) 	<input type="checkbox"/>	Nippon Yusen Kaisha S.S.Kosu	

[Select All](#) [Unselect All](#) [Export Clusters](#) [Merge Selected & Re-Cluster](#) [Merge Selected & Close](#) [Close](#)

Step 8: Redo of Trim Leading and Trailing White Spaces & Collapse consecutive white spaces

Cleaning Dish.csv file

Column: name

As main target of the U_1 is to identify dish details. Hence cleaning of dish name column is the most import step to make sure we get the correct details about the dish based on the name.

Step 1: Trim Leading and Trailing White Spaces

Step 2: Collapse consecutive white spaces

Step 3: GREL expression to remove special characters

The screenshot shows a 'Custom text transform on column name' dialog. The 'Expression' field contains the GREL code: `value.replace(/;/|\t|\n| |"|\\"/, " ")`. The 'Language' dropdown is set to 'General Refine Expression Language (GREL)'. Below the expression, a preview table shows the transformation of six dish names:

row	value	transformed value
1.	Consomme printaniere royal	Consomme printaniere royal
2.	Chicken gumbo	Chicken gumbo
3.	Tomato aux croutons	Tomato aux croutons
4.	Onion au gratin	Onion au gratin
5.	St. Emilion	St. Emilion
6.	Radishes	Radishes

At the bottom, there are options for handling errors: 'On error' (radio buttons for 'keep original', 'set to blank', or 'store error'), a checkbox for 'Re-transform up to 10 times until no change', and buttons for 'OK' and 'Cancel'.

Step 4: Cluster and Edit Column – fingerprint

Note: this was one of the longest steps due to very high cluster count of 40K+. Hence this cleaning was performed in multiple steps by picking 4-5K clusters at a time.

Cluster & Edit column "name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision Keying Function fingerprint 40874 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? <input type="checkbox"/>	New Cell Value
44	47	<ul style="list-style-type: none"> • EGGS poached on toast (2) (3 rows) • EGGS, poached on toast (2) (2 rows) • (2) Poached Eggs on Toast (1 rows) • 2 (Eggs) Poached on Toast (1 rows) • 2 Eggs Poached on Toast (1 rows) • 2 Eggs, Poached on Toast (1 rows) • 2 Eggs, Poached, on Toast (1 rows) • 2 Eggs, poached on toast (1 rows) • 2 Poached (Eggs) on Toast (1 rows) • 2 Poached (eggs) on Toast (1 rows) • 2 Poached Eggs on Toast (1 rows) • 2 Poached Eggs on toast (1 rows) • 2 Poached on Toast, Eggs (1 rows) • 2 eggs poached on toast (1 rows) • 2 eggs, poached on toast (1 rows) • 2 eggs, poached, on toast (1 rows) • EGGS Poached (2) on Toast (1 rows) • EGGS poached on toast 2 (1 rows) • EGGS, poached on toast 2 (1 rows) • Eggs (2) poached on toast (1 rows) • Eggs (2), Poached on Toast (1 rows) • Eggs - Poached on Toast (2) (1 rows) • Eggs Poached (2), on Toast (1 rows) • Eggs poached on toast (2) (1 rows) • Eggs poached on toast 2 (1 rows) • Eggs poached, on toast (2) (1 rows) • Eggs, 2 Poached on Toast (1 rows) 	<input type="checkbox"/> EGGS poached on toast (2)	

Choices in Cluster

2 — 44

Rows in Cluster

2 — 49

Average Length of Choices

0 — 660

Length Variance of Choices

0 — 34

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Cluster & Edit column "name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision Keying Function fingerprint 40874 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? <input type="checkbox"/>	New Cell Value
33	36	<ul style="list-style-type: none"> • Potatoes, French Fried (1 rows) • Potatoes- French Fried (1 rows) • Potatoes-- French fried (1 rows) • french fried potatoes (1 rows) • potatoes, french fried (1 rows) 	<input type="checkbox"/> Hashed Browned Potatoes	Hashed Browned Potatoes

Choices in Cluster

2 — 44

Rows in Cluster

2 — 49

Average Length of Choices

0 — 660

Length Variance of Choices

0 — 34

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Cluster & Edit column "name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision ▾ Keying Function fingerprint ▾ 36573 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? <input type="checkbox"/>	New Cell Value
5	5	<ul style="list-style-type: none"> Home-Made Mince Pie, Hot or Cold (1 rows) Home-made Mince Pie, Hot or Cold (1 rows) Home-made mince pie, hot or cold (1 rows) Homemade Mince Pie, (Hot or Cold) (1 rows) home-made mince pie, hot or cold (1 rows) 	<input type="checkbox"/>	Home-Made Mince Pie, Hot or C
5	5	<ul style="list-style-type: none"> Baked Pork and Beans, with Boston Brown Bread (1 rows) Baked Pork with Boston Beans and Brown Bread (1 rows) Boston Baked Beans and Brown Bread with Pork (1 rows) Boston Baked Beans and Pork with brown bread (1 rows) Boston baked pork and beans with brown bread (1 rows) 	<input type="checkbox"/>	Baked Pork and Beans, with Bc
5	5	<ul style="list-style-type: none"> French Fresh Strawberry Ice Cream (1 rows) French ice cream, fresh strawberry (1 rows) Fresh French Strawberry Ice Cream (1 rows) Fresh Strawberry French Ice Cream (1 rows) Fresh Strawberry french Ice Cream (1 rows) 	<input type="checkbox"/>	French Fresh Strawberry Ice Cr
5	5	<ul style="list-style-type: none"> Assorted Hors d'Oeuvres (If Taken as a Course (1 rows) 	<input type="checkbox"/>	Assorted Hors d'Oeuvres (If Ta

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Choices in Cluster

2 — 5

Rows in Cluster

2 — 12

Average Length of Choices

1 — 50

Length Variance of Choices

0 — 21

Step 5: Cluster and Edit Column – ngram-fingerprint

Cluster & Edit column "name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision ▾ Keying Function ngram-fingerprint ▾ Ngram Size 2 6171 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? <input type="checkbox"/>	New Cell Value
14	35	<ul style="list-style-type: none"> (11 rows) B (4 rows) (3) (3 rows) (1) (2 rows) (2) (2 rows) A (2 rows) C (2 rows) P (2 rows) S?? (2 rows) 9 (1 rows) M (1 rows) Y (1 rows) i. (1 rows) o (1 rows) 	<input type="checkbox"/>	Unknown
6	20	<ul style="list-style-type: none"> Blue Points half shell (12 rows) Blue-Points, half shell (3 rows) Blue Points, Half-Shell (2 rows) Blue Points-Half-Shell (1 rows) Blue points,half shell (1 rows) BluePoints-half shell (1 rows) 	<input type="checkbox"/>	Blue Points half shell
5	16	<ul style="list-style-type: none"> Brussels Sprouts (8 rows) Brussel Sprouts (5 rows) Brussels Sprouts (1 rows) 	<input type="checkbox"/>	Brussels Sprouts

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Choices in Cluster

2 — 14

Rows in Cluster

2 — 48

Average Length of Choices

0 — 510

Length Variance of Choices

0 — 12

Cluster & Edit column "name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method **key collision** Keying Function **ngram-fingerprint** Ngram Size **2** 6171 clusters found

4	10	<ul style="list-style-type: none"> • Home-made Sausages (1 rows) • Homemade Sausage (1 rows) • Blue point oysters on half shell (5 rows) • Blueprint Oysters (on half shell) (3 rows) • Blue Point Oysters on Half-Shell (1 rows) • Blue Point Oysters-on Half Shell (1 rows) 	<input type="checkbox"/> Blue point oysters on half shell
4	4	<ul style="list-style-type: none"> • N.Y.N.H. & H. R. R. Marine Disc. No 1 (1 rows) • N.Y.N.H. & H.R.R. Marine Disc. No 1 (1 rows) • N.Y.N.H.&H.R.R. Marine Disc. No 1 (1 rows) • N.Y.N.H.&H.R.R.Marine Disc. No 1 (1 rows) 	<input type="checkbox"/> N.Y.N.H. & H. R. R. Marine Disc.
4	5	<ul style="list-style-type: none"> • Old-fashioned Strawberry Short Cake, Whipped Cream (2 rows) • Old Fashioned Strawberry Short Cake, Whipped Cream (1 rows) • Old Fashioned Strawberry Shortcake, Whipped Cream (1 rows) • Old-Fashioned Strawberry Shortcake, Whipped Cream (1 rows) 	<input type="checkbox"/> Old-fashioned Strawberry Short
4	10	<ul style="list-style-type: none"> • FROZEN EGGNOD (4 rows) • FROZEN EGG NOGG (3 rows) • Frozen Egg Nog (2 rows) • Frozen Egg-Nogg (1 rows) 	<input type="checkbox"/> FROZEN EGGNOD

Choices in Cluster
2 — 14

Rows in Cluster
2 — 48

Average Length of Choices
0 — 510

Length Variance of Choices
0 — 12

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Step 6: Redo of Trim Leading and Trailing White Spaces & Collapse consecutive white spaces

Column: times_appeared

This was one of the additional cleaning steps.

values less than 0 converted to 0

27 values changed

Custom text transform on column times_appeared

Expression Language **General Refine Expression Language (GREL)** No syntax error.

Preview	History	Starred	Help
44. 458	458		
45. 776	776		
46. 3	3		
47. 124	124		
48. 64	64		
49. 3	3		
50. 157	157		

On error keep original set to blank store error Re-transform up to 10 times until no change

OK Cancel

Data quality changes

Summary of changes

Following is a summary of the changes done and the impact of each change in terms of number of cells got changed based on each step.

Column Name	Step Performed	Cells Cleaned
sponsor	Trim Leading and Trailing White Spaces	0
sponsor	Collapse consecutive white spaces	127
sponsor	GREL expression to remove special characters	880
sponsor	Cluster and Edit Column - fingerprint	5243
sponsor	ngram-fingerprint	1844
sponsor	Trim Leading and Trailing White Spaces	1
sponsor	Collapse consecutive white spaces	1
event	Trim Leading and Trailing White Spaces	0
event	Collapse consecutive white spaces	6
event	GREL expression to remove special characters	342
event	GREL expression to replace ‘?’	85
event	Trim Leading and Trailing White Spaces	19
event	Removing additional ‘?’	24
event	Collapse consecutive white spaces	0
event	Cluster and Edit Column - fingerprint	5460
event	ngram-fingerprint	2311
occasion	Trim Leading and Trailing White Spaces	0
occasion	Collapse consecutive white spaces	3
occasion	GREL expression to remove special characters	2465
occasion	GREL expression to replace ‘?’	171
occasion	Trim Leading and Trailing White Spaces	50
occasion	Removing additional ‘?’	2

occasion	Collapse consecutive white spaces	0
occasion	Cluster and Edit Column - fingerprint	915
occasion	ngram-fingerprint	272
date	Trim Leading and Trailing White Spaces	0
date	Collapse consecutive white spaces	0
date	GREL expression to remove special characters	4
location	Trim Leading and Trailing White Spaces	0
location	Collapse consecutive white spaces	555
location	GREL expression to remove special characters	755
location	GREL expression to replace ‘?’	110
location	Removing additional ‘?’	14
location	Trim Leading and Trailing White Spaces	51
location	Collapse consecutive white spaces	1
location	Cluster and Edit Column - fingerprint	3561
location	ngram-fingerprint	1333
Name (Dish)	Trim Leading and Trailing White Spaces	0
Name (Dish)	Collapse consecutive white spaces	6582
Name (Dish)	GREL expression to remove special characters	10677
Name (Dish)	Trim Leading and Trailing White Spaces	364
Name (Dish)	Collapse consecutive white spaces	71
Name (Dish)	Cluster and Edit Column - fingerprint This was done in 12 steps, as cluster count was very high 40,874	120,392
Name (Dish)	ngram-fingerprint	25341

Before and After

This tables shows how a sample set of data looked before and after changes for different columns cleaned.

Notes: This was not possible to create facet for dish name column due to high cluster count of more than 40K+

Column	Before Changes	After Changes
sponsor	<p>x - sponsor change</p> <p>6354 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> '95 LAW OF COLUMBIAN UNIVERSITY 1 '97S CLASS DINNER 1 'POSSUM CLUB 1 "Bergen" 5 "Conte Biancamano" 1 "Ex Libris" 1 "Hamburg" 2 "Let 'Er Buck" 1 "Paris" 1 "Victoria Luise" 1 (?COLONIAL HOTEL?) 1 (238 EIGHT AVENUE) 1 (ABBAS II HILMI KHEDIVE OF EGYPT) 1 (ADMIRAL LACOMBE?) 1 (ALTA VISTA HOTEL) 4 	<p>x - sponsor change</p> <p>5851 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> '95 LAW OF COLUMBIAN UNIVERSITY 1 '97S CLASS DINNER 1 'POSSUM CLUB 1 12TH REGIMENT INFANTRY 1 14th Regiment Armory 1 15 OFFICERS OF THE PENN RR 1 18 West Thirty Third Street 1 18TH CLUB 1 19th Hole 1 edit include 2 Cultures 1 21 Club 4 21 Club, Inc. 1 22ND REGIMENT A.G.S.A.Y. 1 238 EIGHT AVENUE 2 2601 Parkway 1 26th Anniversary Of New Haven
event	<p>x - event change</p> <p>1769 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> ? 18 '96S DECENTNIAL 1 'FAREWELL TO OLD DELMONICO'S 1 "ANNUAL OUTBREAK" 1 "CLIO" MENU 1 "COMING OF AGE" BANQUET 1 "DEED TO H.M. HOWARD" 1 (?ALUMNI BANQUET?) 1 (?DINNER?) 1 (?LUNCH?) 7 (CARTE DU JOUR) 1 (CELEBRATION OF WASHINGTON'S BIRTHDAY?) 1 (DAILY MENU) 2 (DINNER DANCE) 1 	<p>x - event change</p> <p>1640 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> 'FAREWELL TO OLD DELMONICO'S 1 100TH ANNIVERSARY OF BIRTH OF DANIEL WEBSTER 1 100TH DINNER 1 101ST ANNIVERSARY DINNER 1 edit include 102ND REGULAR MEETING 1 107TH ANNIVESARY DINNER 1 108TH ANNIVERSARY DINNER 1 109TH ANNIVERSARY 1 10NTH ANNIVERSARY DINNER 1 10NTH REUNION DINNER 1 10TH ANNIVERSARY 1 10TH ANNIVERSARY DINNER 1 10TH ANNUAL BANQUET 1 10TH ANNUAL BANQUET OF NATIONAL ASSOCIATION OF CLOTHIERS 1

occasion	<p>occasion change</p> <p>424 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> ; 1 ? 49 "CONTINENTAL DINNER"EN ROUTE; 1 (ANNIV CELEBRATION) 1 (ANNIV?); 1 (ANNIV); 1 (COMMEMOATIVE?) 2 (COMMEMORATIVE); 1 (COMPL;FOR MAHARAJA SCINDIYA) 1 (DAILY MENU) 1 (MEETING) 1 (OTHER - REUNION?); 1 (POSSIBLY A PRIVATELY HOSTED PARTY); 1 	<p>occasion change</p> <p>327 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> OTHER ANNIV 1 OTHER COMMEMORATIVE 1 OTHER COMMERCIAL - MAYBE 1 OTHER SOC - MAYBE 1 10NTH REUNION 1 113 ANNIVERSARY 1 13TH ANNIVERSARY 1 159NTH ANNIVERSARY DINNER-WASHINGTON'S BIRTHDAY 1 21ST ANNUAL MEETING 1 22ND 1 25TH ANNIVERSARY AS ORGANIST & CHOIR MASTER 1 27NTH ANNIVERSARY 1 2ND ANNUAL GAME DINNER 1 2ND GENERAL COVRT 1
location	<p>location change</p> <p>6280 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> ? 48 ? (J B) 1 ? Club 1 ? Hotel 1 '95 Law Of Columbian University 1 '97s Class Dinner 1 'Possum Club 1 "Bergen" 5 "Conte Biancamano" 1 "Ex Libris" 1 "Hamburg" 2 "Let 'Er Buck" 1 "Paris" 1 "Victoria Luise" 1 (Fifth Avenue Hotel?) 1 (Harvard University?) New American House, Boston, Ma 1 [New York Central System] 1 [Not given] 2 	<p>location change</p> <p>5968 choices Sort by: name count Cluster</p> <ul style="list-style-type: none"> - TENTATIVE 48 '95 Law Of Columbian University 1 '97s Class Dinner 1 'Possum Club 1 12th Regiment Infantry 1 14th Regiment Armory 1 15 Officers Of The Penn Rr 1 18 West Thirty Third Street 1 18th Club 1 19th Hole 1 2 Cultures 1 21 Club 4 21 Club, Inc. 1 22nd Regiment A.G.S.A.Y. 1 238 Eight Avenue 2 2601 Parkway 1 26th Anniversary Of New Haven Lodge No. 25 B.P.O.C. 1 27th Assembly District Republican

ICV Queries

Various ICV queries were performed using SQLite in Jupyter notebook to validate the dataset and verify the results of cleaning and use case U_1

These are described below with screenshots

1. When a dish was appeared on which menu, sponsor, event, place (limiting to 10 counts).
 Example dish 'Fried Eggs'. **Use case U1**

```
In [12]: %%sql
select dish.name, menu.sponsor, menu.event, menu.place, dish.first_appeared, menu.date
from menu inner join menupage on menu.id = menupage.menu_id
inner join menuitem on menupage.id = menuitem.menu_page_id
inner join dish on menuitem.dish_id = dish.id
where dish.name = 'Fried Eggs'
limit 10

* sqlite:///menus.db
Done.
```

Out[12]:	name	sponsor	event	place	first_appeared	date
	Fried Eggs	CUNARD LINE	BREAKFAST	SS ETRURIA	1856	1900-02-21
	Fried Eggs				1856	1900-07-08
	Fried Eggs				1856	1900-07-08
	Fried Eggs				1856	1900-07-01
	Fried Eggs	BATTERY PARK HOTEL	BREAKFAST	ASHEVILLE,NC	1856	1900-02-21
	Fried Eggs				1856	1900-06-01
	Fried Eggs				1856	1900-07-08
	Fried Eggs	CANADIAN PACIFIC RAILWAY COMPANY	BREAKFAST	RMS EMPRESS OF CHINA	1856	1900-04-23
	Fried Eggs				1856	1901-11-07
	Fried Eggs				1856	1901-10-05

2. When a dish was **first** appeared on which menu, sponsor, event, place (limiting to 10 counts).
 Example dish 'French Fries'. **Use case U1**

```
In [13]: %%sql
select dish.name, menu.sponsor, menu.event, menu.place, menu.date, dish.first_appeared
from menu inner join menupage on menu.id = menupage.menu_id
inner join menuitem on menupage.id = menuitem.menu_page_id
inner join dish on menuitem.dish_id = dish.id
where dish.name = 'French Fries'
and menu.date is not ''
order by menu.date
limit 1

* sqlite:///menus.db
Done.
```

Out[13]:	name	sponsor	event	place	date	first_appeared
	French Fries	Hotel Worthy	DINNER	SPRINGFIELD, MASS.	1901-07-05	1901

3. When a dish was **last** appeared on which menu, sponsor, event, place (limiting to 10 counts).
 Example dish 'Fried Eggs'. **Use case U1**

```
In [14]: %%sql
select dish.name, menu.sponsor, menu.event, menu.place, menu.date, dish.last_appeared
from menu inner join menupage on menu.id = menupage.menu_id
inner join menuitem on menupage.id = menuitem.menu_page_id
inner join dish on menuitem.dish_id = dish.id
where dish.name = 'Fried Eggs'
and menu.date is not ''
order by menu.date desc
limit 1

* sqlite:///menus.db
Done.
```

Out[14]:	name	sponsor	event	place	date	last_appeared
	Fried Eggs	Holland America Cruises			1987-01-01	1987

Additional verifications

4. Verify uniqueness of id field in files. No data output with these queries means, all of these IDs were unique in each file.

Menu Table

```
In [6]: %%sql
select *
from menu
where id in (select id as c
from menu
group by id
having count(id) > 1 )
* sqlite:///menus.db
Done.

Out[6]: id name sponsor event venue place physical_description occasion notes call_number keywords language date location location_type currency curre
```

Dish Table

```
In [7]: %%sql
select *
from dish
where id in (select id as c
from dish
group by id
having count(id) > 1 )
* sqlite:///menus.db
Done.

Out[7]: id name description menus_appeared times_appeared first_appeared last_appeared lowest_price highest_price
```

Menupage Table

```
In [8]: %%sql
select *
from menupage
where id in (select id as c
from menupage
group by id
having count(id) > 1 )
* sqlite:///menus.db
Done.
```

Menu Item Table

```
In [17]: %%sql
select *
from menuitem
where id in (select id as c
from menuitem
group by id
having count(id) > 1 )
* sqlite:///menus.db
Done.
```

5. Each Menu item should have a single menu page id - Verified

```
In [20]: %%sql

select id, menu_page_id
from menuitem
where menu_page_id in (select id as c
from menuitem
group by id
having count(menu_page_id) > 1 )

* sqlite:///menus.db
Done.

Out[20]: id menu_page_id
```

6. Dish Count less than menu page count – 150 instances found against this constraint, but this could be due to other details and pictures in the menu. Hence no data was changed for this. But it was an interesting discovery

Ideally, dish count should be more than menu page count. But there are some instances where this is not correct. This could be due to other details and pictures in the menu

```
In [38]: %%sql

select count(*)
from menu
where cast(page_count as integer) > cast(dish_count as integer)

* sqlite:///menus.db
Done.

Out[38]: count(*)
150
```

7. Top 5 dishes with most appearances

```
In [9]: %%sql

select name, count(name)
from dish
group by name
order by count(name) desc
limit 5

* sqlite:///menus.db
Done.

Out[9]:
```

	name	count(name)
	Fried sweet potatoes	49
	EGGS poached on toast (2)	48
	Cold Roast Beef	47
	Grape Fruit, Half	45
	G H Mumm's Extra Dry	45

8. Top 5 dishes appearing on the most number of menus

```
In [10]: %%sql
select id, name, menus_appeared
from dish
order by menus_appeared desc
limit 5
* sqlite:///menus.db
Done.
```

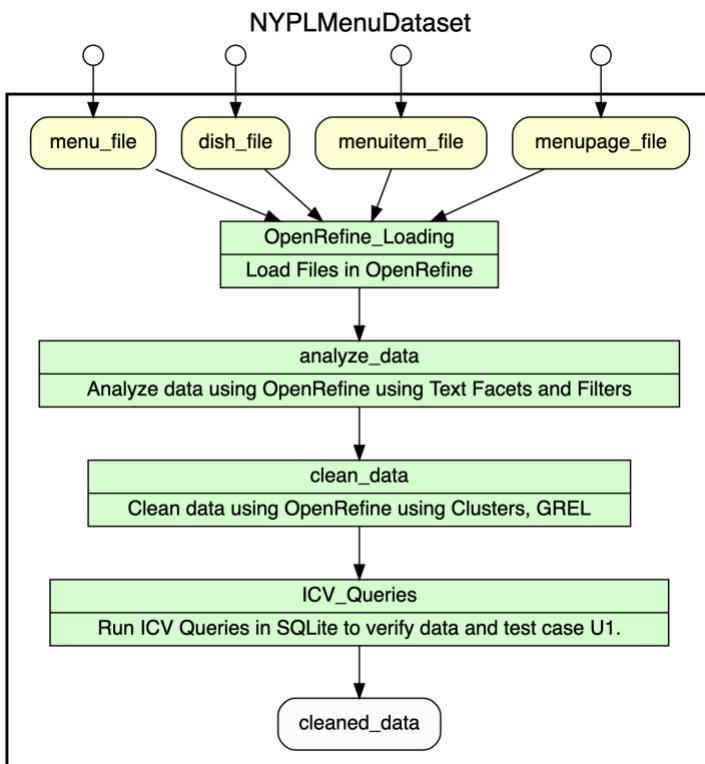
```
Out[10]:
      id          name  menus_appeared
150  Little Neck clams            995
1034 Anchovies on toast             99
1047   *Roast Capon                99
1372    Creme Yvette                99
2798  Pork Tenderloin                99
```

Workflow Model

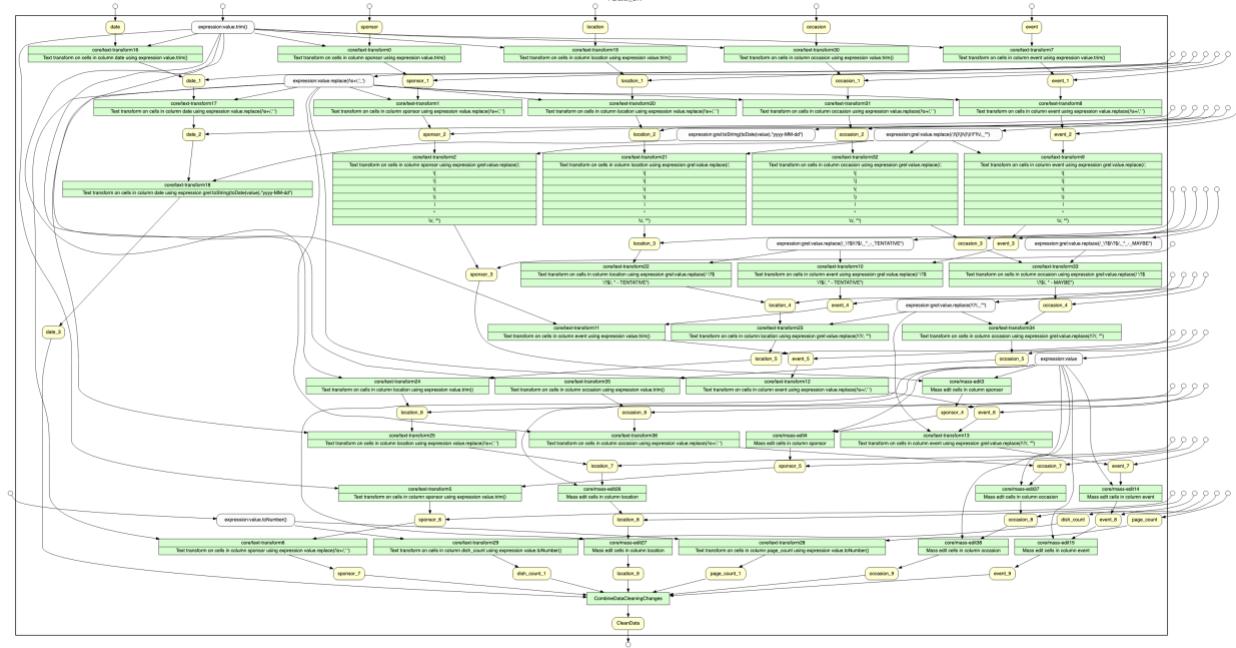
Workflow W0:

Overall Workflow of the process used in this exercise.

- OpenRefine was used as a tool to clean the data like remove special characters using GREL, Clustering, format conversion and other data cleaning steps
- SQLite was used to run ICV queries using Jupyter notebook

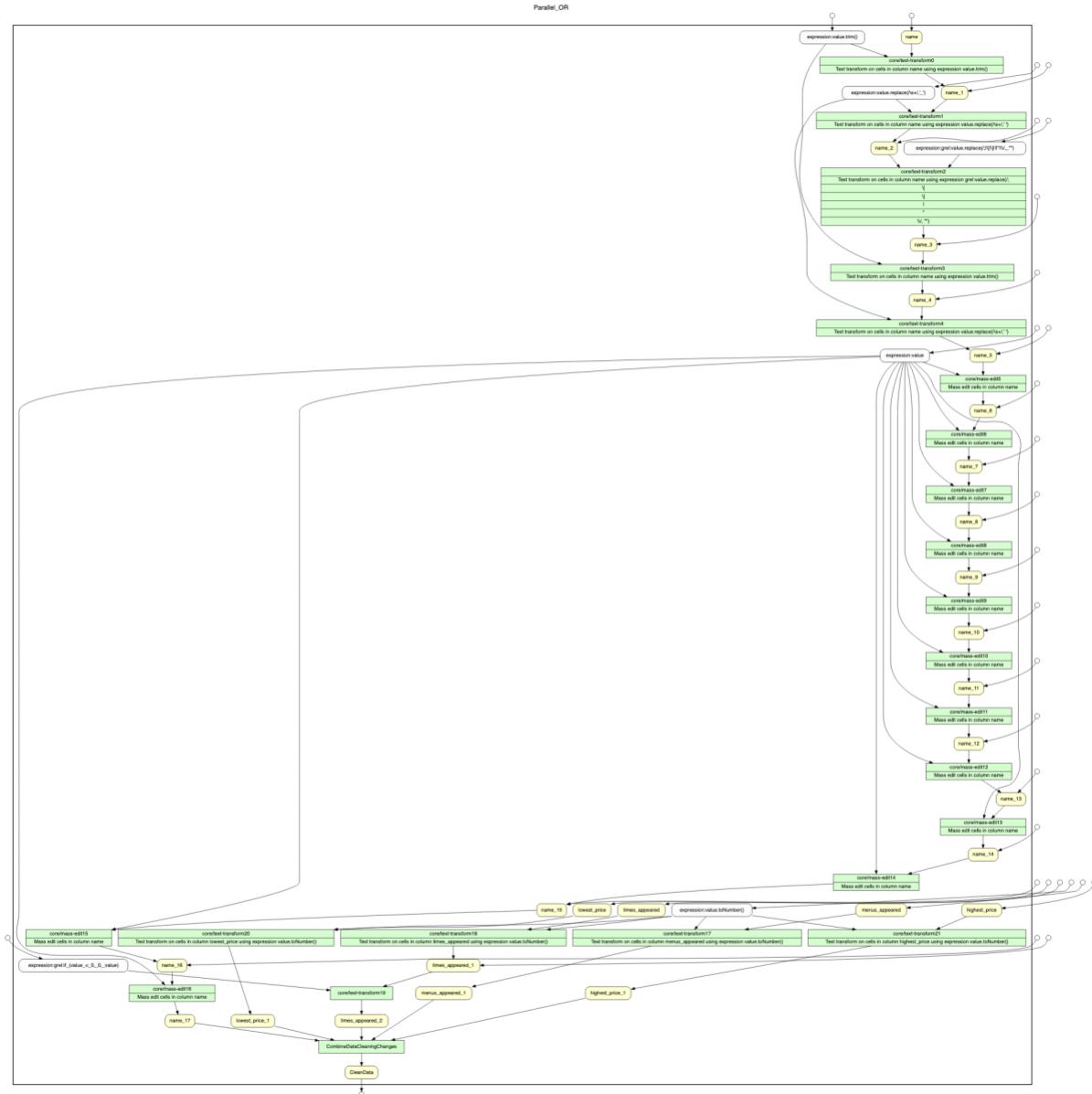


Workflow W1: Cleaning Menu.csv file



Note: Due to complex workflow model all the elements may not be visible in this report. So, a png file is separately added as well along with the submission

Workflow W2: Cleaning Dish.csv file



Note: Due to complex workflow model all the elements may not be visible in this report. So, a png file is separately added as well along with the submission

Conclusions & Summary

This data cleaning activity was quite fun and great learning activity. Working with OpenRefine, SQLite, using SQLite on Jupyter notebook, using YesWorkflow was a great experience. One of the big challenges of the activity was cleaning a huge dataset, specially Dish.csv file. Number of clusters were more than 40K and OpenRefine just kept running for hours. Then I decided to filter the clusters and perform the Steps by dividing it into 12 steps. As a team of one individual, I did perform all the activities in this project.

Although two files MenuItem and MenuPage are very less mentioned in this report. The reason is almost not data cleaning was needed for these two. But these two were key file to get the linking between a Dish and its respective Menu.

Overall, as a summary I was able to clean more than 190K cells including multiple steps and able to achieve target to find dish details like on which Menu dish first or last appeared.