2022

CS513: THEORY & PRACTICE OF DATA CLEANING - FINAL PROJECT

NAVDEEP PARMAR - NAVDEEP2@ILLINOIS.EDU
PRASHANTH GIRIYAPPA SHANKARAPPA - PG14@ILLINOIS.EDU
SHUBHENDU BHASKAR - SB59@ILLINOIS.EDU

CONTENTS

1 INTRODUCTION

This document describes reiterates the use cases, overall data cleaning process, tools used, data cleaning steps and retrospection from executing this data cleaning project.

All the artifacts related to this project are placed here.

2 CHOSEN DATASET

For the data cleaning project, we are choosing "The New York Public Library" menu's dataset(http://menus.nypl.org/data).

3 USE CASES- REVISITED

Scenario: "The New York Public Library" menu's dataset has menus from restaurants that have been transcribed over the years. The restaurants, their menus and dishes have seen changes through the years. For restaurant market research analysis, we would like to gather information about these menus and associated dishes, events, and occasions where these dishes were served and associated prices. With this information we plan to build repository/website which would serve information from the past and more recent years, which can also provide statistics and options to pick from. Examples of questions are stated below in the Target/Main use case.

3.1 TARGET/MAIN USE CASE(U1):

We revised Use case (U1) from our initial assessment because they were returning empty sets.

a. What dishes were served as "Breakfast" meal for "Easter" event at a "Commercial" venue organized in descending order of price in US dollars?
a. What are the different styles of eggs served for Breakfast that were on the menu in 1900s and 2000s at different venues?
b. What are the different varieties of wines were on the menu in 1900s and 2000s which were served at different venues for "Easter"?
c. Between 1990 and 2000, what are the restaurants that served traditional dishes starting from 1900s in Lunch or Dinner?

Initial Assessment(U1)

A. Find the highest price of dishes (dish.csv) along with their names that were available at the dinner (menu.csv -event) at "Government" (menu.csv - venue) on the occasion of "Anniversary" (menu.csv - Occasion) irrespective of the currency.
B. What are the different varieties of wines were in the menu in 1900s vs 2000s in each of the Venue setting (Commercial, Government, SOC etc.) on an Occasion of "Anniversary"?
C. What are the different styles of eggs as a breakfast option were in the menu in 1900s vs. 2000s in each of the Venue setting (Commercial, Government, SOC etc.)?

D. In 2000s, what were some of the places that you could go for Lunch and Dinner that served traditional and new dishes?

3.2 NO CLEANING REQUIRED USE CASE (U0):

We revised U0 from our initial assessment. The below mentioned Initial assessment returned different set of records on cleaned and raw data, which defied the purpose of U0. The Revised Use case(U0) returns same set of records with and without cleaning operations.

U0: What are the top 10 high prices of menu items across all menus?

Initial assessment U0: How many dishes are similar or repeated in the Dishes Dataset?

3.3 NOT SUFFICIENT CLEANING USE CASE (U2):

How many menus were non-English and were Handwritten?

4 DATA CLEANING WORKFLOW

From our initial analysis we had established that Data cleaning workflow will have multiple phases – Data Cleaning with Open Refine, Intermediate step using tools like Python/Pandas and final cleaning stage using SQL. In each stage we performed certain operations identified for cleaning the dataset.

A link to Overall Workflow is available here. OpenRefine offers several common transformation and clustering methods that were used to clean datasets while visualizing the results. After dataset is cleaned by OpenRefine, we used Pandas dataframe to perform additional cleaning finally ending in SQL. Using database tables, we were able to establish relationships between different dataset which also helped us remove sematic integrity constraint violations.

Using OpenRefine Using OpenRefine helped us visualize dataset and correct **syntactic violations**. We were able to remove special characters, trim leading or trailing whitespaces, transform fields with numbers or dates to visualize clusters, group and merge values that are similar. These operations helped us cluster relevant field values together to answer questions that use case proposes. From Menu dataset we need information about Name of restaurant, Event, Venue, Occasion, location, date, and currency. From Dish dataset we are interested in name, lowest_price and highest_price fields since we are trying to answer highest prices.

Using Python/Pandas Cleaned datasets from OpenRefine will be used by Python and Pandas package to remove columns that are not required by our use case. Examples of these include but are not limited to fields like physical_description, notes, language etc. From Dish dataset fields that are not relevant are menus_appeared, times_appeared etc. We will also remove fields/columns that are empty from these datasets. At the end of this phase of cleaning we will produce cleaned ".csv" files with only columns that are relevant to the use case.

Using relational database and SQL In this phase we will create tables based on relational schema and insert records into the tables. We will review relational tables for **Semantic violations, duplicate entries,**

and null values. This will help us further remove inconsistencies in the datasets. Using SQL, we will query the database and present results for our use case.

5 Tools Used

- **OpenRefine** for general transformations and clustering datasets
- **Python and Pandas** to perform additional cleaning steps
- **SQL** to check for semantic violations and duplicates
- **"Pandas profiling"** python package for profiling datasets
- **"YesWorkflow Web Application"** to create Overall Yesworkflow
- **"Openrefine to Yesworkflow model tool" (OR2YWTool)** to convert ". json" history to ".pdf" and ". yw" files

6 Data Cleaning steps

There are 4 datasets that are available – Menu, Dish, MenuItem and MenuPage. We established relations between these datasets in our initial assessment and schema diagram is available here. We dedicated much of our time cleaning Menu and Dish datasets since these datasets contain information required for our main use case. MenuItem and MenuPage datasets are used to establish relationship with Menu and Dish datasets since they have keys that connect them.

Below are the common data cleaning operations that we have used in each dataset:

- Trimmed whitespaces from the values
- Collapsed consecutive whitespaces
- Removal of Unicode characters using GREL -regular expressions
- Remove columns with no values present
- Convert the text values to Title Case
- Clustered the values wherever applicable

6.1 Data Cleaning using OpenRefine

6.1.1 Menu Dataset

Menu dataset consists of features that can be associated with a menu. Visualizations from the cleaning operations for Menu dataset are stored here.

DATASET FEATURES	
Total number of records	17545
Total number of columns	20
Number of columns that are blank	3

6.1.1.1 Cleaning Operations: We performed a total of 145 steps to clean Menu dataset. Many columns were cleaned and some of the transformations were performed multiple times.

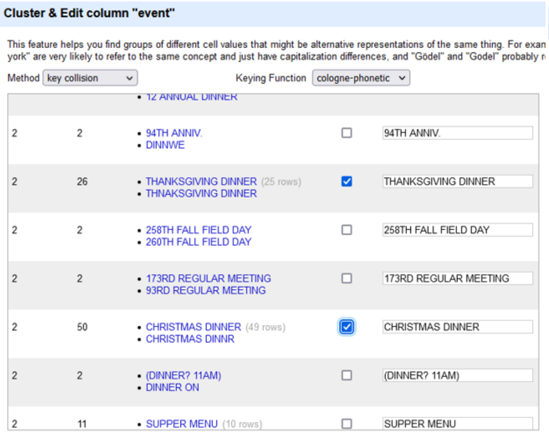
a. Extensive cleaning on fields - **name, event, venue, place, occasion, location and status**. We carefully chose fields that must be merged based on our use case.
   i. Common transformations:
      1. Trim white spaces from field values
      2. Collapse consecutive whitespaces
      3. Convert text values to Title case
   ii. Clustering performed
      1. Key Collison with Fingerprint
      2. Key Collison with ngram-fingerprint ngram=2
      3. Other keying functions were used but were selectively chosen so that they wouldn't cluster different locations together. For instance, using "metaphone3" to group different Alumini associations from different regions together was not desirable for our use case.
   iii. Regular expressions
      1. Remove [?()\/] symbols
      2. Replace symbols from the beginning of values

b. **Currency** - Trim and Change to Title case. Total cells changed 2,575

c. Number of changes performed - Location 11837, Event 21524, Venue 15298, Occasion 1257, Value 3036, Place 11468, Name 2736.

d. Screenshots from cluster and merge

Menu - Sponsor and Event fields: Although we used different key functions, our objective was to selectively choose values to merge.

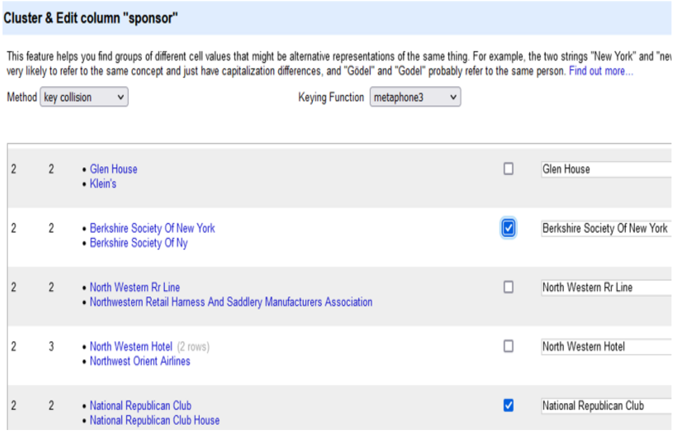


Cluster & Edit column "event"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For exan
york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably r

Method: key collision   Keying Function: cologne-phonetic

		• 12 ANNUAL DINNER		
2	2	• 94TH ANNIV. • DINNWE	□	94TH ANNIV.
2	26	• THANKSGIVING DINNER (25 rows) • THNAKSGIVING DINNER	☑	THANKSGIVING DINNER
2	2	• 258TH FALL FIELD DAY • 260TH FALL FIELD DAY	□	258TH FALL FIELD DAY
2	2	• 173RD REGULAR MEETING • 93RD REGULAR MEETING	□	173RD REGULAR MEETING
2	50	• CHRISTMAS DINNER (49 rows) • CHRISTMAS DINNR	☑	CHRISTMAS DINNER
2	2	• (DINNER? 11AM) • DINNER ON	□	(DINNER? 11AM)
2	11	• SUPPER MENU (10 rows)	□	SUPPER MENU



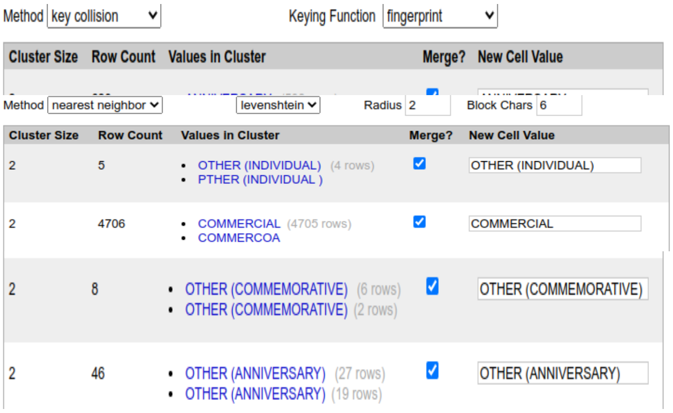
Cluster & Edit column "sponsor"

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 "ne
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: metaphone3

2	2	• Glen House • Klein's	□	Glen House
2	2	• Berkshire Society Of New York • Berkshire Society Of Ny	☑	Berkshire Society Of New York
2	2	• North Western Rr Line • Northwestern Retail Harness And Saddlery Manufacturers Association	□	North Western Rr Line
2	3	• North Western Hotel (2 rows) • Northwest Orient Airlines	□	North Western Hotel
2	2	• National Republican Club • National Republican Club House	☑	National Republican Club

Menu – Occasion: Our objective with Clustering occasion field was normalize as much as possible and reduce number of clusters. Here you can see that we disregarded suggested corrections.

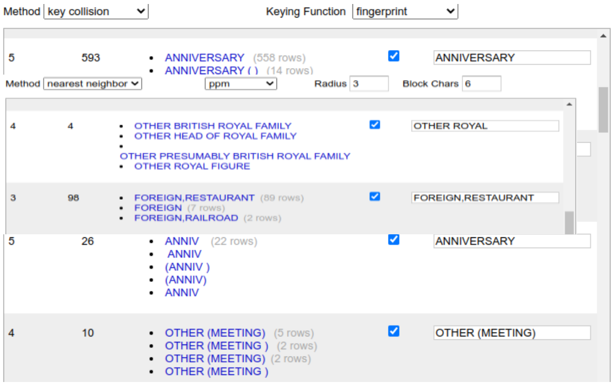


Method key collision Keying Function fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value

Method nearest neighbor levenshtein Radius 2 Block Chars 6

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	5	OTHER (INDIVIDUAL) (4 rows) PTHER (INDIVIDUAL)	☑	OTHER (INDIVIDUAL)
2	4706	COMMERCIAL (4705 rows) COMMERCOA	☑	COMMERCIAL
2	8	OTHER (COMMEMORATIVE) (6 rows) OTHER (COMMEMORATIVE) (2 rows)	☑	OTHER (COMMEMORATIVE)
2	46	OTHER (ANNIVERSARY) (27 rows) OTHER (ANNIVERSARY) (19 rows)	☑	OTHER (ANNIVERSARY)



Method key collision Keying Function fingerprint

5	593	ANNIVERSARY (558 rows) ANNIVERSARY ( ) (14 rows)	☑	ANNIVERSARY

Method nearest neighbor ppm Radius 3 Block Chars 6

4	4	OTHER BRITISH ROYAL FAMILY OTHER HEAD OF ROYAL FAMILY OTHER PRESUMABLY BRITISH ROYAL FAMILY OTHER ROYAL FIGURE	☑	OTHER ROYAL
3	98	FOREIGN,RESTAURANT (89 rows) FOREIGN (7 rows) FOREIGN,RAILROAD (2 rows)	☑	FOREIGN,RESTAURANT

5	26	ANNIV (22 rows) ANNIV (ANNIV) (ANNIV) ANNIV	☑	ANNIVERSARY
4	10	OTHER (MEETING) (5 rows) OTHER (MEETING) (2 rows) OTHER (MEETING) (2 rows) OTHER (MEETING)	☑	OTHER (MEETING)

Menu – Venue: We used different key functions, and this is a field that is required by use case.

6.1.2 Dish Dataset

Dish dataset consists of features that can be associated with a dish. This dataset had fewer columns but larger number of records. Visualization from the cleaning operations for Menu dataset are stored here.

DATASET FEATURES	
Total number of records	423397
Total number of columns	9
Number of columns that are blank	1

6.1.2.1 Cleaning Operations:

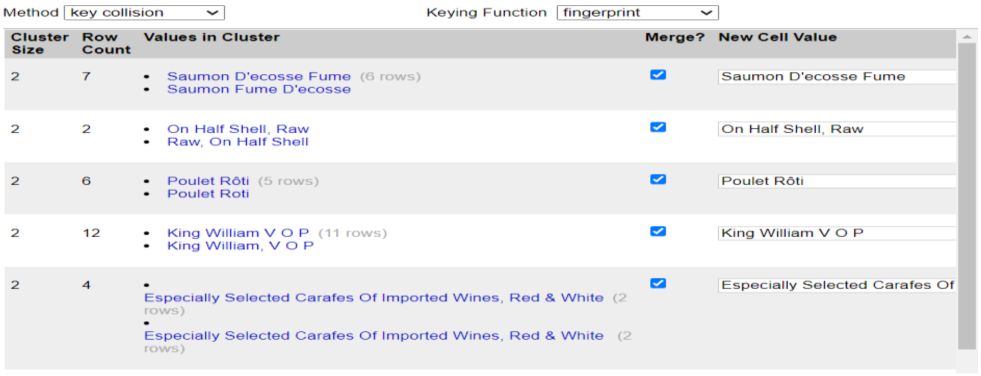
a. Columns - id, menus_appeared, times_appeared, first_appeared, last_appeared, lowest_price, highest_price
   i. Trim white spaces from field values
   ii. Collapsed consecutive whitespaces
   iii. Since this column has numeric values, convert the values of the column to “Number”

b. Column - Name
   i. Common transformations performed
      - Trim white spaces from field values: This was performed multiple times before and after clustering – Cells affected 9045 before clustering and 18598 after clustering
      - Collapse consecutive whitespaces - Cells affected 35791
      - Convert text values to Title case - Cells affected 271393

   ii. Regular expressions
      - Replace some special character “? ";.*\” - Cells affected 43415
      - Replace special characters at the beginning of dish name – Cells affected 1617
      - Replace special characters at the end of dish name - Cells affected 2158

iii. Cluster and Merge - This step was challenging because of the number of clusters that were generated using different methods. Large number of clusters were generated by OpenRefine, so we had to Cluster and merge in stages.

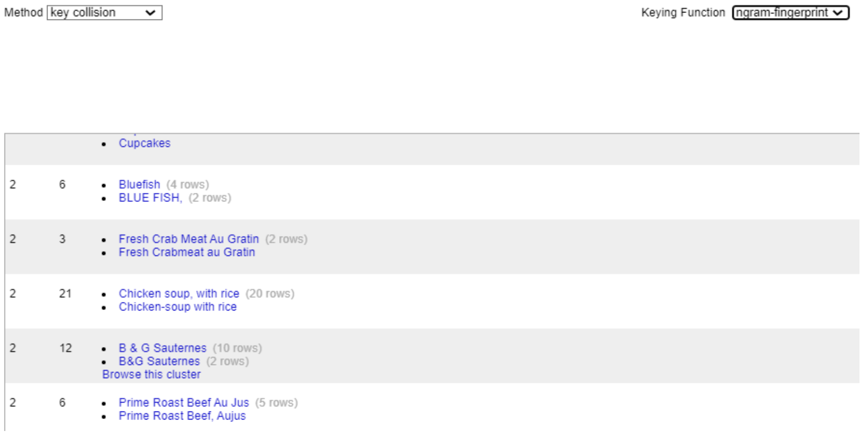
- **Key Collison with Fingerprint** - Our use case refers to dishes and we require dish names to be clustered. Cells affected 107,966



Method key collision Keying Function fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	7	Saumon D'ecosse Fume (6 rows) Saumon Fume D'ecosse	☑	Saumon D'ecosse Fume
2	2	On Half Shell, Raw Raw, On Half Shell	☑	On Half Shell, Raw
2	6	Poulet Rôti (5 rows) Poulet Roti	☑	Poulet Rôti
2	12	King William V O P (11 rows) King William, V O P	☑	King William V O P
2	4	Especially Selected Carafes Of Imported Wines, Red & White (2 rows) Especially Selected Carafes Of Imported Wines, Red & White (2 rows)	☑	Especially Selected Carafes Of

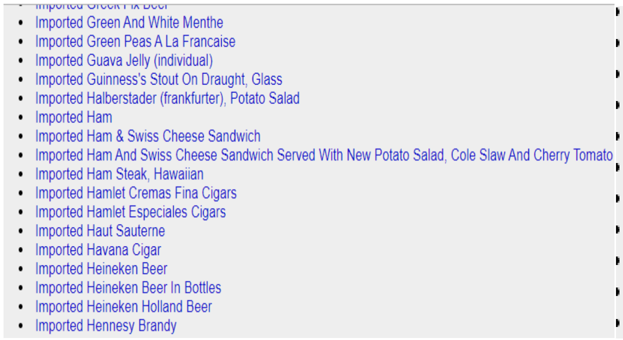
- **Key Collison with ngram-fingerprint with 2-gram** – Cells affected 15511



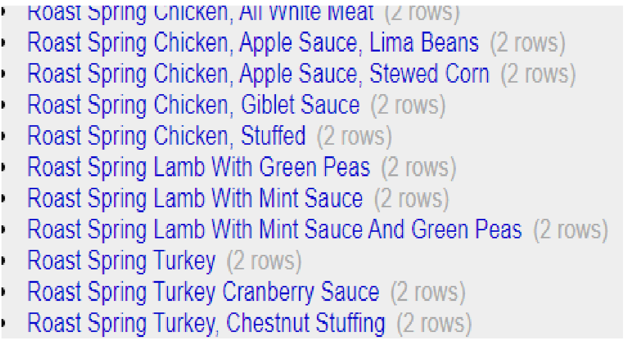
Method key collision Keying Function ngram-fingerprint

		Cupcakes
2	6	Bluefish (4 rows) BLUE FISH, (2 rows)
2	3	Fresh Crab Meat Au Gratin (2 rows) Fresh Crabmeat au Gratin
2	21	Chicken soup, with rice (20 rows) Chicken-soup with rice
2	12	B & G Sauternes (10 rows) B&G Sauternes (2 rows) Browse this cluster
2	6	Prime Roast Beef Au Jus (5 rows) Prime Roast Beef, Aujus

- **Other cluster and merge algorithms were NOT used** since it would generalize dish names and we would lose valuable information needed by our use case. For instance, metaphone2 would have combined Roasted Chicken, Lamb, and Turkey together. Another example where all "imported" items are clustered together which is not what we want for our use case.



- Imported Green And White Menthe
- Imported Green Peas A La Francaise
- Imported Guava Jelly (individual)
- Imported Guinness's Stout On Draught, Glass
- Imported Halberstader (frankfurter), Potato Salad
- Imported Ham
- Imported Ham & Swiss Cheese Sandwich
- Imported Ham And Swiss Cheese Sandwich Served With New Potato Salad, Cole Slaw And Cherry Tomato
- Imported Ham Steak, Hawaiian
- Imported Hamlet Cremas Fina Cigars
- Imported Hamlet Especiales Cigars
- Imported Haut Sauterne
- Imported Havana Cigar
- Imported Heineken Beer
- Imported Heineken Beer In Bottles
- Imported Heineken Holland Beer
- Imported Hennesy Brandy



- Roast Spring Chicken, All White Meat (2 rows)
- Roast Spring Chicken, Apple Sauce, Lima Beans (2 rows)
- Roast Spring Chicken, Apple Sauce, Stewed Corn (2 rows)
- Roast Spring Chicken, Giblet Sauce (2 rows)
- Roast Spring Chicken, Stuffed (2 rows)
- Roast Spring Lamb With Green Peas (2 rows)
- Roast Spring Lamb With Mint Sauce (2 rows)
- Roast Spring Lamb With Mint Sauce And Green Peas (2 rows)
- Roast Spring Turkey (2 rows)
- Roast Spring Turkey Cranberry Sauce (2 rows)
- Roast Spring Turkey, Chestnut Stuffing (2 rows)

6.1.3 MenuPage Dataset

MenuPage dataset establishes the relationship of menu and page numbers. We didn’t find many opportunities for cleaning MenuPage dataset as it was mostly clean.

DATASET FEATURES	
Total number of records	66397
Total number of columns	7
Number of columns that are blank	0

6.1.3.1 Cleaning Operations:

a. id, menu_id, page_number, image_id, full_height, full_width:
   i. Trim white spaces from field values
   ii. Collapsed consecutive whitespaces
   iii. Since this column has numeric values, convert the values of the column to “Number

6.1.4 MenuItem Dataset

MenuItem dataset establishes between MenuPage and Dish. We didn’t find many opportunities for cleaning MenuPage dataset.

DATASET FEATURES	
Total number of records	1,095,130
Total number of columns	9
Number of columns that are blank	0

6.1.4.1 Cleaning Operations:

a. id, menu_page_id, price, high_price, dish_id, created_at, updated_at, xpos, ypos:
   i. Trim white spaces from field values
   ii. Collapsed consecutive whitespaces
   iii. Since this column has numeric values, convert the values of the column to “Number

Datasets cleaned using OpenRefine can be accessed here.

6.2 DATA CLEANING USING PYTHON

6.2.1 Menu Dataset

a. "Sponsor” column was removed from our dataset as it has the same values as “Location” column. Location column is more relevant as we see more non-blank values when compared to “Sponsor” column.
b. Removed columns “keywords”, “language”, “location_type” as their values were blank
c. Removed columns “physical_description”, “notes”, “call_number”, “currency_symbol”as they are not relevant to our use cases
d. Removed rows that had status of “under review”. Cells affected - 174

e. We have replaced the blank values of the column “currency” to “Dollar” as we found that the majority blanks values of “currency” belonged to the United States. Cells affected - 11089

f. Date field was converted to date with format “YYYY-mm-dd”, and invalid dates were coerced.

6.2.2 Dish Dataset

a. Removed column “description” as its values were blank.

Datasets cleaned using Python can be accessed here.

6.3 Data Cleaning using SQL

6.3.1 Semantic violations and duplicate records checks

We created database tables and run SQL to check for integrity constraint violations and for additional updates based on established relationships. Queries that were run are available here.

a. Primary key, NOT NULL constraints
    - i. Menu Table Integrity Constraint, Duplicate keys, and NULL value check - Rows cleaned 0
    - ii. Dish Table Integrity Constraint, Duplicate keys, and NULL value check - Rows cleaned 0
    - iv. MenuPage and MenuItem Table Integrity Constraint, Duplicate keys, and NULL value check - Rows cleaned 0

Primary key constraints were not violated in any dataset.

b. Referential Integrity constraints
    - i. Menu and MenuPage tables ON Menu.id - There are 6379 Menu Ids that are missing from MenuPage table which indicates that menu pages were not recorded for these menus. These menus will eventually be excluded from our final query.
    - ii. MenuItem and Dish tables ON Dish Id - There are 5,809 Dish Ids that are not mentioned on MenuItem table. MenuItem connects MenuPage and Dish tables.

c. Other cleaning operations performed
    - i. Menu table
        1. **“name”** column - Replace Blank, NULL and Not Given with “NA” - Rows updated 14,328
        2. **“occasion”** column - Replace Blank, NULL and Not Given with “NA” - Rows updated 13,658
        3. **“event”** column - Replace Blank, NULL and Not Given with “NA” - Rows updated 9,305
        4. **“venue”** column - Replace Blank, NULL and Not Given with “NA” - Rows updated 9,344
        5. **“date”** column - Replace Blank and NULL with “9999-12-31” - Rows updated 579
        6. **“location”** column - Replace Blank and NULL with “NA” - Rows updated 261
    - ii. Dish table
        1. **“location”** column - Replace Blank and NULL with “NA” - Rows updated 45
        2. **“lowest_price”** column - Replace Blank and NULL with 0 - Rows updated 29,100

3. **“highest_price”** column - Replace Blank and NULL with 0 - Rows updated 29,100

iii. MenuItem table

1. **“price”** column - Replace Blank and NULL with 0 - Rows updated 445916

Final cleaned dataset can be accessed here.

7 IMPROVEMENT

We compared improvements of our cleaning operations by running queries against cleaned and original datasets. The SQL for the use cases can be found at here. To summarize, we observed that by cleaning the datasets, the volume of records returned was larger and we also got better results for our queries which seems to indicate that similar values were merged.

7.1.1 Comparison using SQL queries

Use Case (U1)

a. What dishes were served as “Breakfast” meal for “Easter” event at a “Commercial” venue organized in descending order of price in US dollars?

Dataset	Records
Cleaned	168
Raw	0

b. What are the different styles of eggs served for Breakfast that were on the menu in 1900s and 2000s at different venues?

Dataset	Records
Cleaned	796
Raw	161

c. What are the different varieties of wines were on the menu in 1900s and 2000s which were served at different venues for “Easter”?

Dataset	Records
Cleaned	2
Raw	0

d. Between 1990 and 2000, what are the restaurants that served traditional dishes starting from 1900s in Lunch or Dinner?

Dataset	Records
Cleaned	7
Raw	44

In this case the raw datasets returned duplicate records.

Use Case (U0)

a. What are the top 10 high prices of menu items across all menus?
Received same sets of records from cleaned and raw datasets.

Use Case (U2)

a. How many menus were non-English and were Handwritten?
From the dataset it appears that the language and type of menu (handwritten or not), can be derived from the "notes" column, which is a collection of subjective information. Further assessment of "notes" depicts that it does not follow specific order of fields. Therefore, no matter how much we clean the data, we can never be sure that each row will have information pertaining to language or type of menu.

7.1.2 Comparison by Profiling cleaned and original datasets

We profiled the cleaned and original datasets using "Pandas Profiling" package. A detailed report from profiling can be accessed here. Here is our comparison for Menu and Dish datasets.

Menu Dataset statistics - The following table shows the result of our cleaning operations for fields on Menu dataset. It can be observed that number of distinct values has decreased in the cleaned dataset which indicates consolidation of values.

Dataset statistics			
	Menu-Cleaned	Menu-Original	Diff (Original-Cleaned)
Overall			
Number of variables	9	9	
Missing cells (%)	30.40%	37.10%	6.70%
Name			
Distinct	797	797	
Distinct (%)	20.10%	24.90%	4.80%
Missing (%)	82.50%	81.80%	-0.70%
Event			
Distinct	1564	1770	
Distinct (%)	19.40%	21.70%	2.30%
Missing (%)	53.60%	53.50%	-0.10%
Venue			
Distinct	70		
Distinct (%)	0.90%	29.00%	28.10%
Missing (%)	53.80%	53.70%	-0.10%
Occasion			
Distinct	229	423	
Distinct (%)	6.20%	11.20%	5.00%
Missing (%)	78.60%	78.40%	-0.20%

Location			
Distinct	5705	6283	
Distinct (%)	33.30%	35.80%	2.50%
Missing (%)	1.50%	0.00%	-1.50%
Currency			
Distinct	42	42	
Distinct (%)	0.20%	0.70%	0.50%
Missing (%)	0.00%	63.20%	63.20%

Dish Dataset statistics - The following table shows the result of our cleaning operations for fields on Dish dataset. Most of the effort was spent on cleaning "name" field of the dataset. The percentage of number of distinct values has decreased by over 20% for "name" field. Fewer distinct values indicates that similar clusters were merged.

Dataset statistics			
	Dish-Cleaned	**Dish-Original**	**Diff (Original-Cleaned)**
Overall			
Number of variables	6	6	
Missing cells (%)	2.30%	2.30%	0.00%
Name			
Distinct	336939	423363	
Distinct (%)	79.60%	99.90%	20.30%
Missing (%)	0.00%	0.00%	0.00%

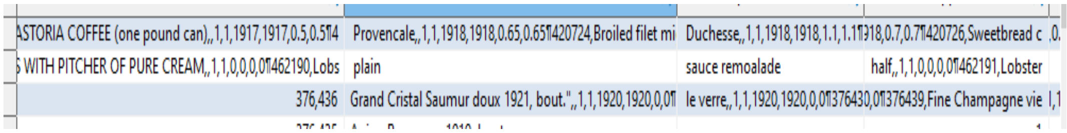
8 RETROSPECTION

8.1 FINDINGS

- In OpenRefine, when selecting Key function, changing default values of parameters like n-gram size etc., can uncover additional clusters.
- Some OpenRefine clustering algorithms can be too aggressive you will lose valuable information if chosen. One example where we went with many options was for Menu dataset's occasion field. For this field we wanted to group as many values together as possible. For Dish dataset's Name field, we didn't want to lose information contained in dish name, so our clustering was more relaxed.
- For our main use case(U1), we are querying dish dataset using "wine" keyword as a filter. In our analysis we found that there are different types of wines available that do not have the keyword "wine" in their name.

8.2 CHALLENGES

- Different cluster key functions can cluster values in different ways. We could not always go with cluster suggestions from OpenRefine tool. Some values had to be chosen. Some of the clusters didn't always make sense for our use case.
- Working with large files in OpenRefine is a challenge, for instance, we had issues importing "MenuItem.csv". The tool got stuck in the hung state causing delays in cleaning operations.
- If the number of clusters are over 50,000, OpenRefine would not cluster correctly. We encountered this with Dish dataset.
- Cluster and merge would not always make grammatical sense for many values. To better cluster we need better algorithms or the use of Natural Language processing.
- We encountered issues while importing the raw Dish.csv in SQL lite database. Approximately, half the records were only committed to the table in the import from the csv. This restricted us in accurately comparing (before and after results from the query of our uses cases) between raw dataset and cleaned dataset.



ASTORIA COFFEE (one pound can),,1,1,1917,1917,0.5,0.5¶4	Provencale,,1,1,1918,1918,0.65,0.65¶420724,Broiled filet mi	Duchesse,,1,1,1918,1918,1.1,1.1¶	318,0.7,0.7¶420726,Sweetbread c	,0.
5 WITH PITCHER OF PURE CREAM,,1,1,0,0,0,0¶462190,Lobs	plain	sauce remoalade	half,,1,1,0,0,0,0¶462191,Lobster	
376,436	Grand Cristal Saumur doux 1921, bout.",,1,1,1920,1920,0,0¶	le verre,,1,1,1920,1920,0,0¶37643	0,0¶376439,Fine Champagne vie	l,1

9 REFERENCES

- What's on the menu? - http://menus.nypl.org
- Openrefine to Yesworkflow model tool - OR2YWTool Link
- Yes workflow editor service Link
- Pandas profiling package Link
- Handbook of Data Quality by Shazia Sadiq