CS513: THEORY & PRACTICE OF DATA CLEANING

Final Project - Phase I

Authors:

Ji Ma (Jay): jima2@illinois.edu

Yutong Wang (Andy): yutong9@illinois.edu

Minjie Fu (Jane): minjief2@illinois.edu

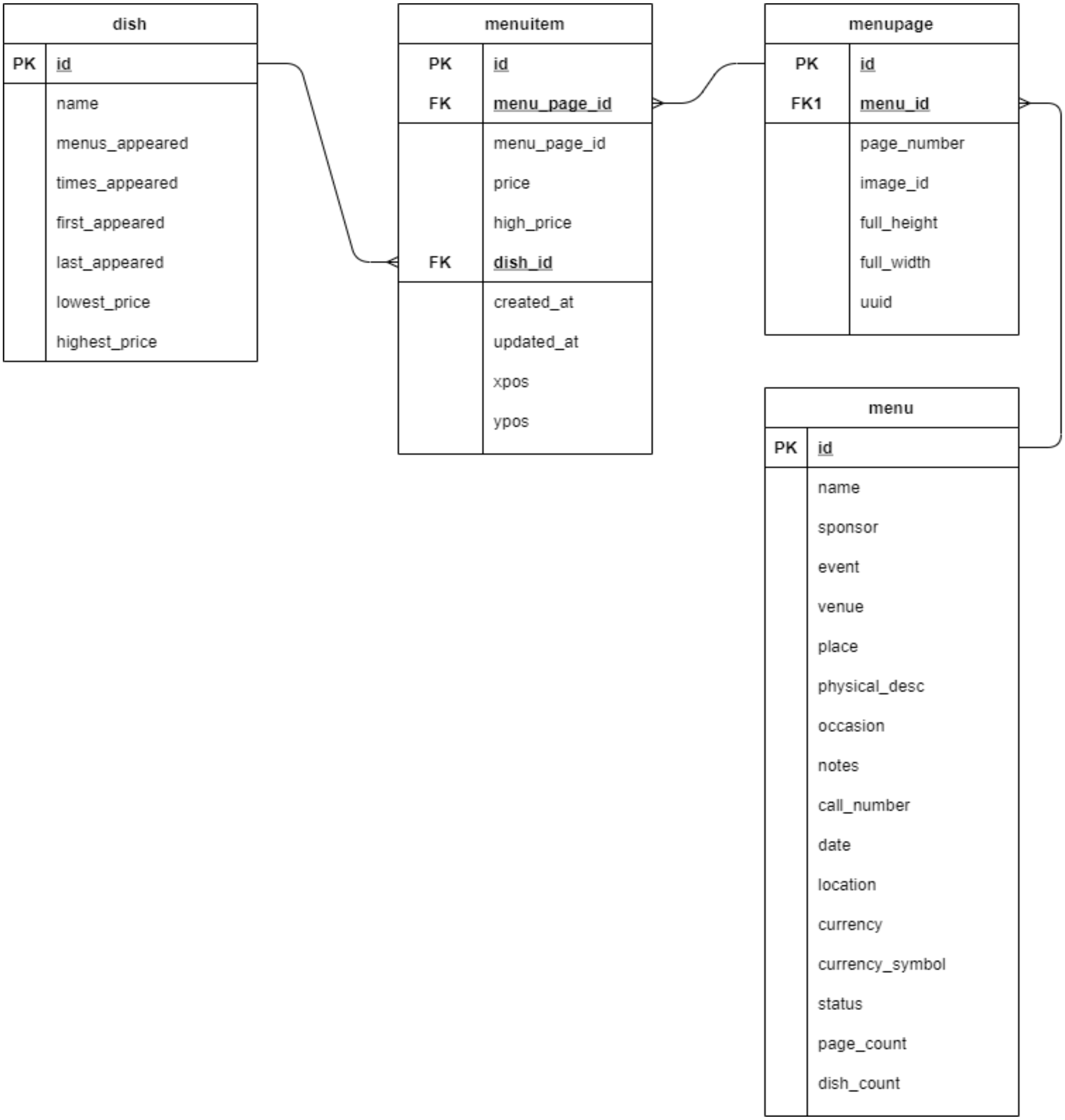
The purpose of this document is to propose an end-to-end data cleaning workflow, in practice of the data cleaning and provenance tools and techniques introduced in course CS513.

1 Database Selected

The database we used is NYPL-menus. It is a project named “What’s on the Menu” NYPL (New York Public Library) labs launched in late April 2011. The project aimed at scanning and transcribing all the 45,000 menus contained in the New York Public Library’s menu collection dating from the 1840s to the present. Till now, a quarter of them have been digitized and collected in this database.

2 Database Description

The database contains 4 tables: Menu, MenuPage, MenuItem, and Dish. The database schema and the relationships between tables are shown in the following ER diagram.



1) Menu

The table “Menu” holds general information regarding the restaurants and the menus they offer. The attribute “id” is the unique identifier for each menu and will be used as the primary key. Other information includes the name of the restaurant, the sponsor, the location, the occasion, the physical description of the menu, number of menu pages, number of dishes on the menu, etc. The table’s schema is shown as follows:

Table: menu

```
CREATE TABLE menu (
id BIGINT PRIMARY KEY
NOT NULL,
name TEXT,
sponsor TEXT,
event TEXT,
venue TEXT,
place TEXT,
physical_desc TEXT,
occasion TEXT,
notes TEXT,
call_number TEXT,
date DATE,
location TEXT,
currency TEXT,
currency_symbol TEXT,
status TEXT,
page_count INTEGER,
dish_count INTEGER
);
```

2) MenuPage

The table “MenuPage” contains information regarding each page of the menu. Each page is identified by a unique id. The attribute “menu_id” shows on which menu the page appears. It is used as the foreign key to link the table “MenuPage” and the table “Menu”. Other information includes the height and width of the page, the page number, the scanned image id, etc. The table’s schema is shown as follows:

Table: menupage

```
CREATE TABLE menupage (
id BIGINT PRIMARY KEY
NOT NULL,
menu_id BIGINT REFERENCES menu (id),
page_number INTEGER,
image_id INTEGER,
full_height INTEGER,
full_width INTEGER,
uuid TEXT
);
```

3) MenuItem

The table “MenuItem” contains information regarding the items shown on the menu. Each item is identified by a unique id. The attribute “menu_page_id” lists the id of the menu page on which the item appears. This attribute can be used as the foreign key to link the table “MenuItem” and the table “Menupage”. Also, the attribute “dish_id” lists the id of the dish that the item is assigned to, which can be used as the foreign key to link the table “MenuItem” and the table “Dish”.

Other information includes the price of the item, the time it was created, and other meta information such as x pos, y pos, which is the position that the item is located on the scanned image. The table’s schema is shown as follows:

```
Table: menuitem
CREATE TABLE menuitem (
id BIGINT PRIMARY KEY
NOT NULL,
menu_page_id BIGINT REFERENCES menupage (id),
price DOUBLE,
high_price DOUBLE,
dish_id BIGINT REFERENCES dish (id),
created_at DATETIME,
updated_at DATETIME,
xpos DOUBLE,
ypos DOUBLE
);
```

4) Dish

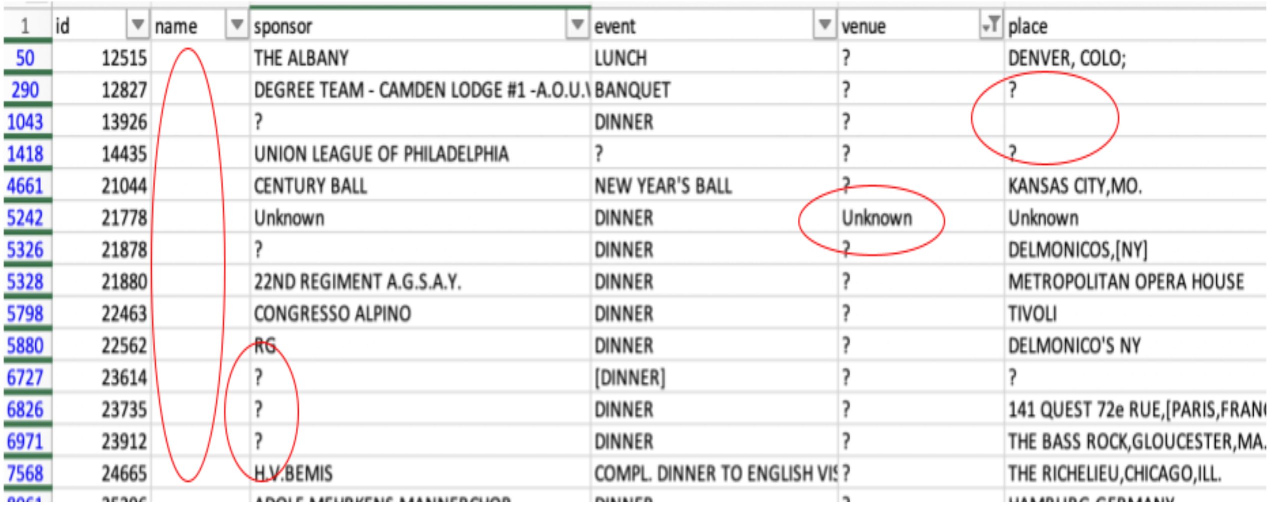
The table “Dish” contains information regarding each dish on the menu. The attribute “id” is the unique identifier for each dish and will be used as the primary key. Other information includes the name of the dish, the number of times it appears on the menus, the first and last year it appears, its lowest and highest prices, etc. The table’s schema is shown as follows:

```
Table: dish
CREATE TABLE dish (
id BIGINT PRIMARY KEY
NOT NULL,
name TEXT,
menus_appeared INTEGER,
times_appeared INTEGER,
first_appeared DATETIME,
last_appeared DATETIME,
lowest_price DOUBLE,
highest_price DOUBLE
);
```

3 Quality issues

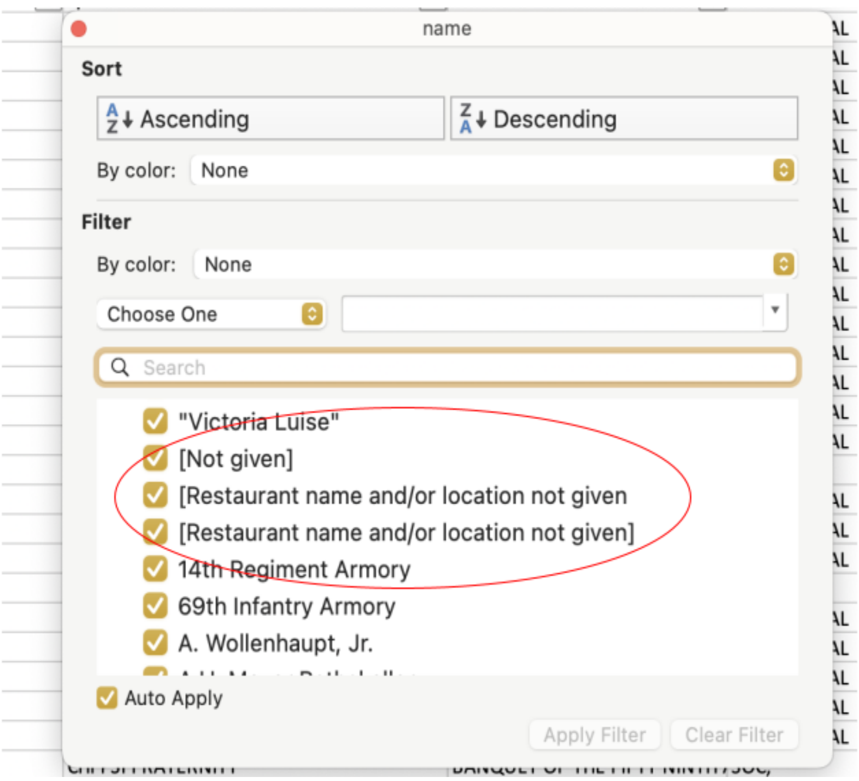
We took an initial inspection of the dataset by using the Filter Function in Excel. We had to admit that the data quality is poor. We spotted and listed some obvious data quality problems as follows:

1) Missing Values and Inconsistent Handling Methods



1	id	name	sponsor	event	venue	place
50	12515		THE ALBANY	LUNCH	?	DENVER, COLO;
290	12827		DEGREE TEAM - CAMDEN LODGE #1 -A.O.U.\	BANQUET	?	?
1043	13926		?	DINNER	?	
1418	14435		UNION LEAGUE OF PHILADELPHIA	?	?	?
4661	21044		CENTURY BALL	NEW YEAR'S BALL	?	KANSAS CITY,MO.
5242	21778		Unknown	DINNER	Unknown	Unknown
5326	21878		?	DINNER	?	DELMONICOS,[NY]
5328	21880		22ND REGIMENT A.G.S.A.Y.	DINNER	?	METROPOLITAN OPERA HOUSE
5798	22463		CONGRESSO ALPINO	DINNER	?	TIVOLI
5880	22562		RG	DINNER	?	DELMONICO'S NY
6727	23614		?	[DINNER]	?	?
6826	23735		?	DINNER	?	141 QUEST 72e RUE,[PARIS,FRAN
6971	23912		?	DINNER	?	THE BASS ROCK,GLOUCESTER,MA.
7568	24665		H.V.BEMIS	COMPL. DINNER TO ENGLISH VIS	?	THE RICHELIEU,CHICAGO,ILL.

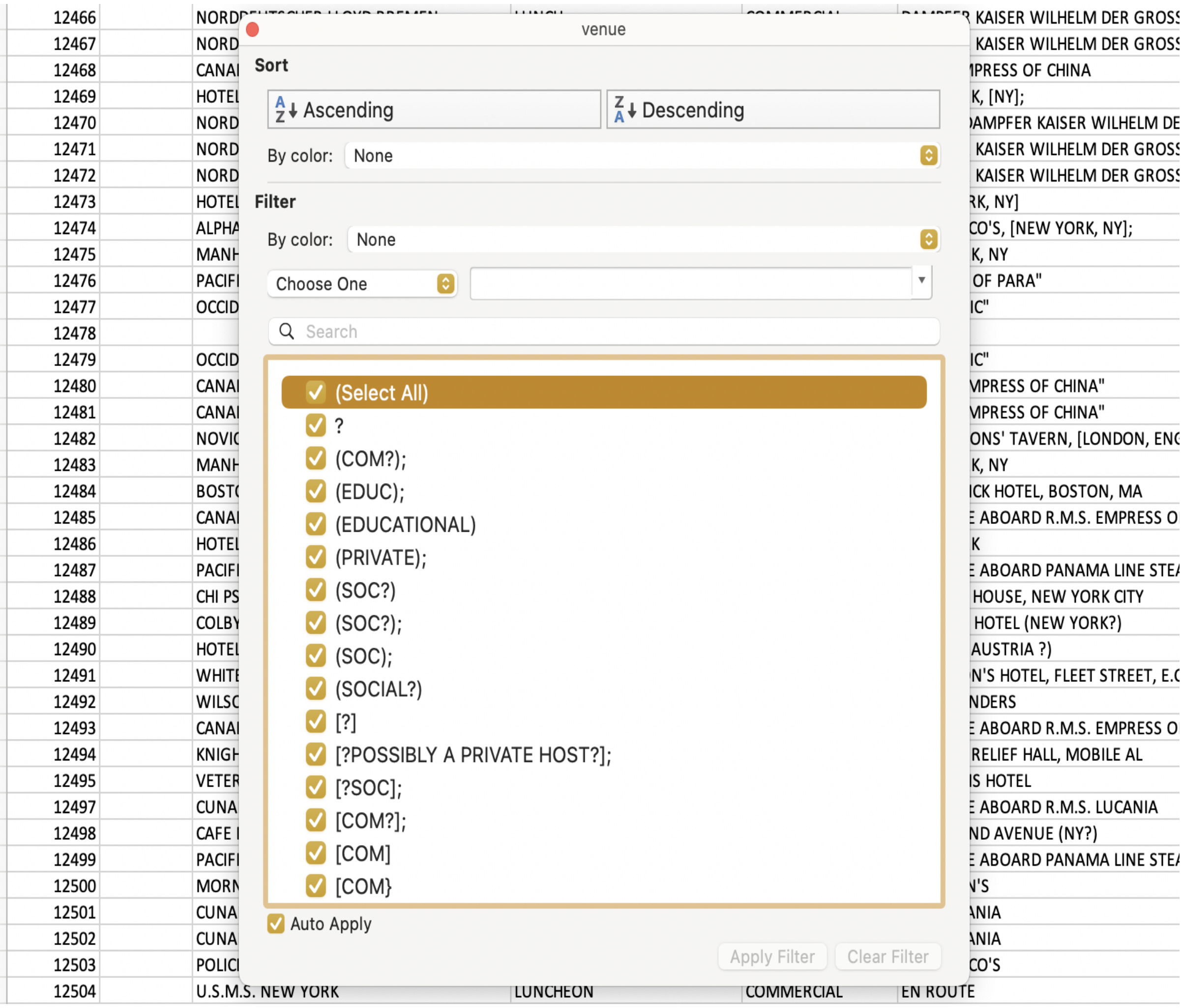
There are a lot of missing values in the dataset, and different methods are used to deal with these missing values. For example, in the table "Menu", some of the missing values are left blank. Some of them are filled in with "?" or "Unknown". In the column "name", it uses the description of "[Not given]" or "[Restaurant name and/or location not given]" to handle the missing values.



☑ "Victoria Luise"
☑ [Not given]
☑ [Restaurant name and/or location not given
☑ [Restaurant name and/or location not given]
☑ 14th Regiment Armory
☑ 69th Infantry Armory
☑ A. Wollenhaupt, Jr.

2) Leading/Trailing Special Characters

Inconsistent special characters are used at the leading and trailing positions through the whole dataset. Take the column “venue” in the table “Menu” as an example. Some fields are enclosed with special characters such as “()”, “[]”, “;” and “?”, which causes the content inconsistency.



12466	NORDDEUTSCHER LLOYD BREMEN	LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSS
12467	NORD			KAISER WILHELM DER GROSS
12468	CANAI			IPRESS OF CHINA
12469	HOTEL			K, [NY];
12470	NORD			DAMPFER KAISER WILHELM DE
12471	NORD			KAISER WILHELM DER GROSS
12472	NORD			KAISER WILHELM DER GROSS
12473	HOTEL			RK, NY]
12474	ALPHA			CO'S, [NEW YORK, NY];
12475	MANH			K, NY
12476	PACIFI			OF PARA"
12477	OCCID			IC"
12478				
12479	OCCID			IC"
12480	CANAI			MPRESS OF CHINA"
12481	CANAI			MPRESS OF CHINA"
12482	NOVIC			ONS' TAVERN, [LONDON, ENG
12483	MANH			K, NY
12484	BOSTC			ICK HOTEL, BOSTON, MA
12485	CANAI			E ABOARD R.M.S. EMPRESS O
12486	HOTEL			K
12487	PACIFI			E ABOARD PANAMA LINE STEA
12488	CHI PS			HOUSE, NEW YORK CITY
12489	COLBY			HOTEL (NEW YORK?)
12490	HOTEL			AUSTRIA ?)
12491	WHITE			N'S HOTEL, FLEET STREET, E.C
12492	WILSC			NDERS
12493	CANAI			E ABOARD R.M.S. EMPRESS O
12494	KNIGH			RELIEF HALL, MOBILE AL
12495	VETER			IS HOTEL
12497	CUNA			E ABOARD R.M.S. LUCANIA
12498	CAFE I			ND AVENUE (NY?)
12499	PACIFI			E ABOARD PANAMA LINE STEA
12500	MORN			N'S
12501	CUNA			ANIA
12502	CUNA			ANIA
12503	POLICI			CO'S
12504	U.S.M.S. NEW YORK	LUNCHEON	COMMERCIAL	EN ROUTE

venue

Sort

By color: None

Filter

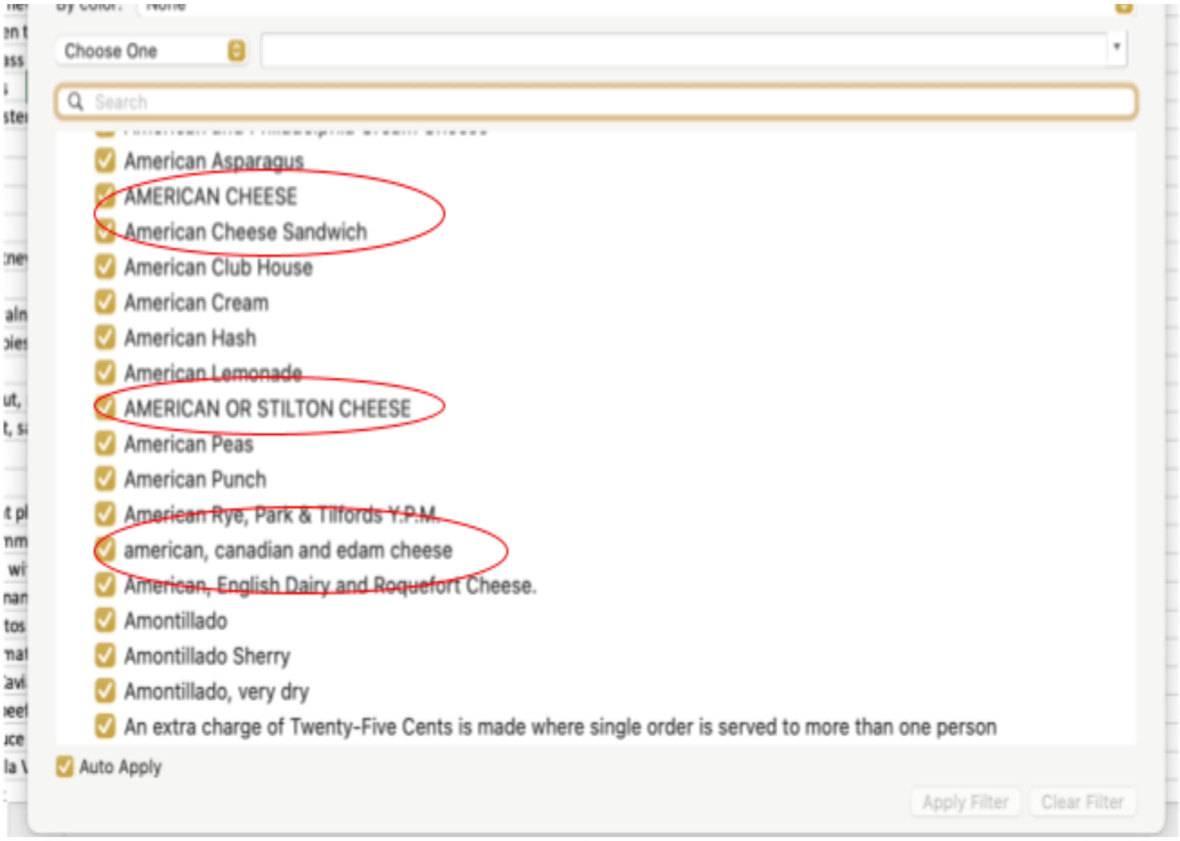
By color: None

☑ (Select All)
☑ ?
☑ (COM?);
☑ (EDUC);
☑ (EDUCATIONAL)
☑ (PRIVATE);
☑ (SOC?)
☑ (SOC?);
☑ (SOC);
☑ (SOCIAL?)
☑ [?]
☑ [?POSSIBLY A PRIVATE HOST?];
☑ [?SOC];
☑ [COM?];
☑ [COM]
☑ [COM}

☑ Auto Apply

3) Inconsistent Upper/Lower Case

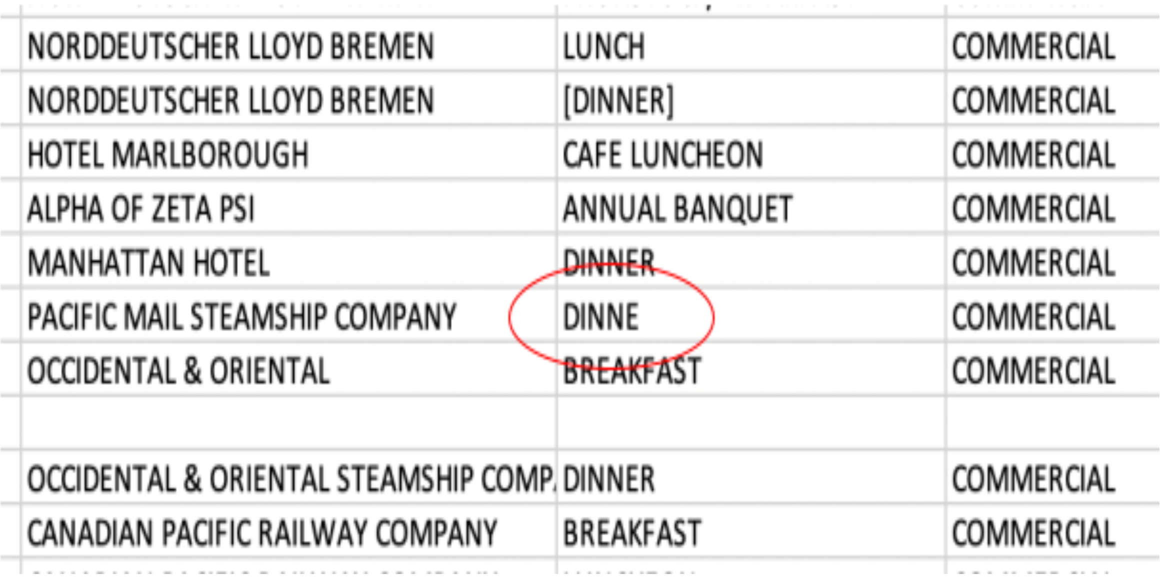
The dataset inconsistently uses uppercase and lowercase letters. Some fields capitalize the first letter, some fields capitalize all the letters, others use all lowercase letters. The following chat highlight an example spotted in the column "name" in the table "Dish".



- ☑ American and Philadelphia Cream Cheese
- ☑ American Asparagus
- ☑ AMERICAN CHEESE
- ☑ American Cheese Sandwich
- ☑ American Club House
- ☑ American Cream
- ☑ American Hash
- ☑ American Lemonade
- ☑ AMERICAN OR STILTON CHEESE
- ☑ American Peas
- ☑ American Punch
- ☑ American Rye, Park & Tilfords Y.P.M.
- ☑ american, canadian and edam cheese
- ☑ American, English Dairy and Roquefort Cheese.
- ☑ Amontillado
- ☑ Amontillado Sherry
- ☑ Amontillado, very dry
- ☑ An extra charge of Twenty-Five Cents is made where single order is served to more than one person

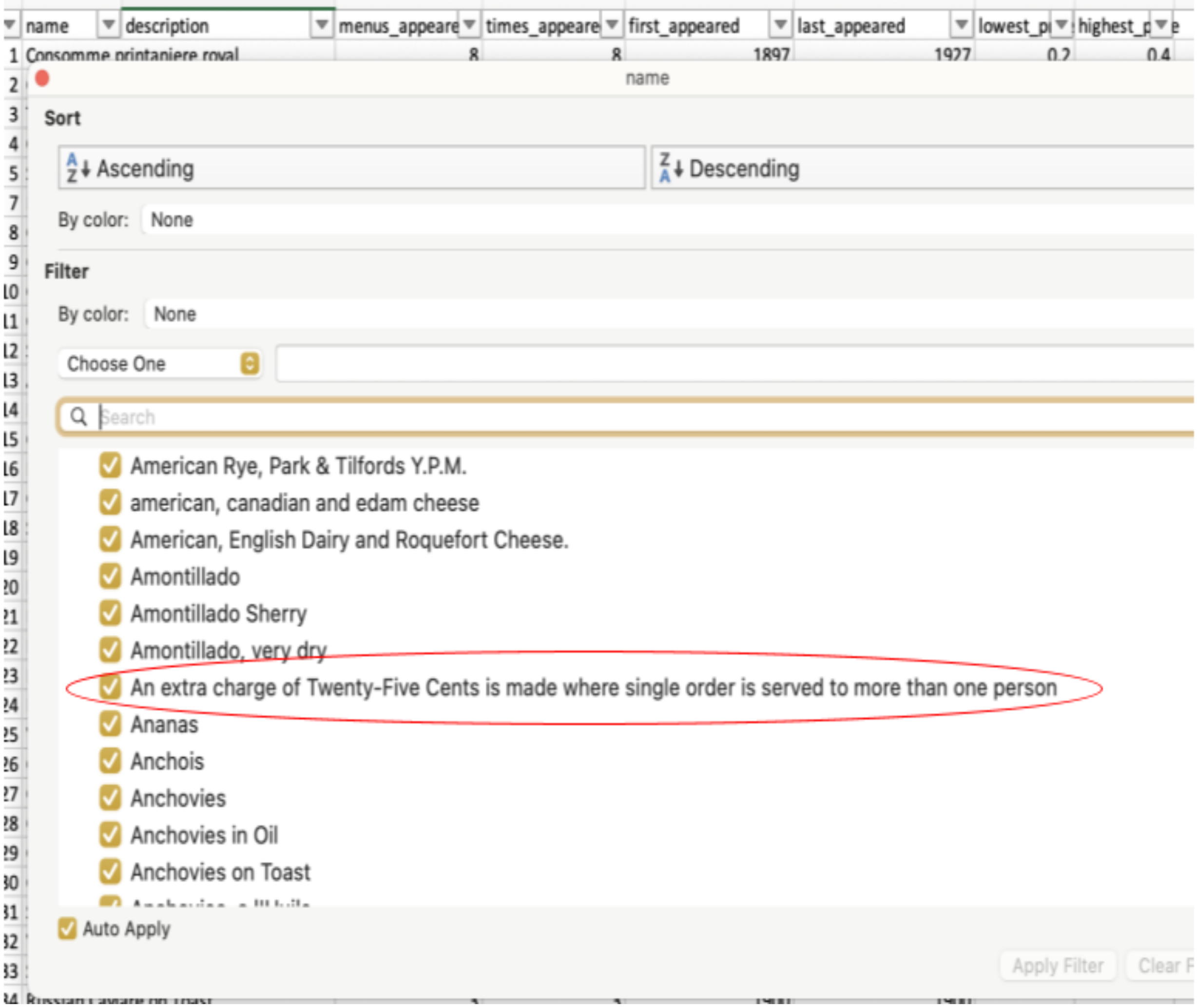
☑ Auto Apply

4) Spelling and Entry Errors



NORDDEUTSCHER LLOYD BREMEN	LUNCH	COMMERCIAL
NORDDEUTSCHER LLOYD BREMEN	[DINNER]	COMMERCIAL
HOTEL MARLBOROUGH	CAFE LUNCHEON	COMMERCIAL
ALPHA OF ZETA PSI	ANNUAL BANQUET	COMMERCIAL
MANHATTAN HOTEL	DINNER	COMMERCIAL
PACIFIC MAIL STEAMSHIP COMPANY	DINNE	COMMERCIAL
OCCIDENTAL & ORIENTAL	BREAKFAST	COMMERCIAL
OCCIDENTAL & ORIENTAL STEAMSHIP COMP.	DINNER	COMMERCIAL
CANADIAN PACIFIC RAILWAY COMPANY	BREAKFAST	COMMERCIAL

The dataset contains a lot of spelling errors. For example, “DINNER” is spelled as “DINNE”. Also, the content of some fields does not match the corresponding column’s header. For example, shown in the following chart, the highlighted content should be filled in the column “description”, not the column “name”. It is obviously an entry error.



	name	description	menus_appeare	times_appeare	first_appeared	last_appeared	lowest_pi	highest_p
1	Consomme printaniere royal		8	8	1897	1927	0.2	0.4

name

Sort

Ascending | Descending

By color: None

Filter

By color: None

Choose One

Search

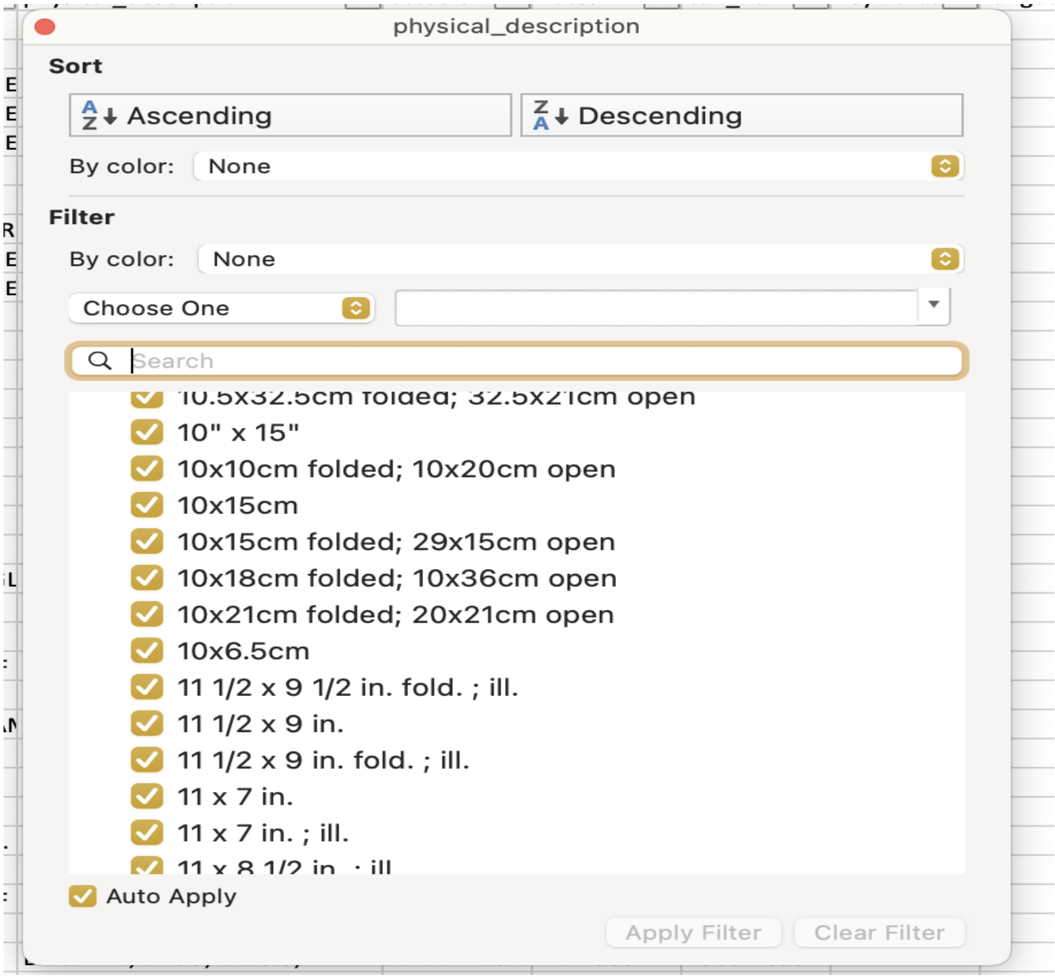
- ☑ American Rye, Park & Tilfords Y.P.M.
- ☑ american, canadian and edam cheese
- ☑ American, English Dairy and Roquefort Cheese.
- ☑ Amontillado
- ☑ Amontillado Sherry
- ☑ Amontillado, very dry
- ☑ An extra charge of Twenty-Five Cents is made where single order is served to more than one person
- ☑ Ananas
- ☑ Anchois
- ☑ Anchovies
- ☑ Anchovies in Oil
- ☑ Anchovies on Toast

☑ Auto Apply

Apply Filter | Clear F

5) Different measurement units and language

The measurement units used through the dataset are not consistent. For example, in the column "physical_description" of the table "Menu", some fields use "cm", while others use "in". The inconsistent measurement units are misleading. Meanwhile, we found some dishes' names use English, while others use Italian, French or other languages, which may also lead to data quality problems.



By color: None

By color: None

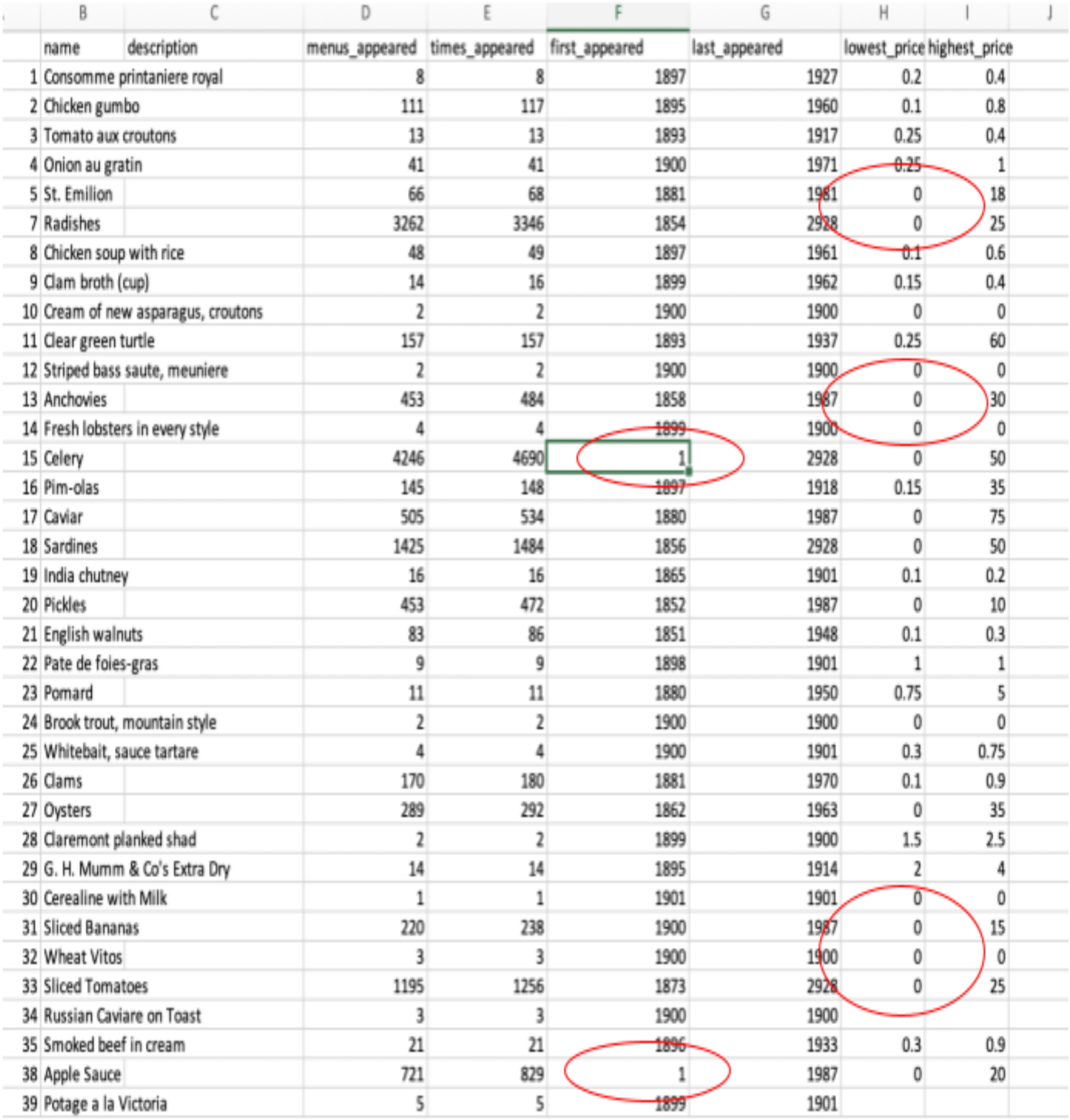
Choose One

☑ 10.5x32.5cm folded; 32.5x21cm open
☑ 10" x 15"
☑ 10x10cm folded; 10x20cm open
☑ 10x15cm
☑ 10x15cm folded; 29x15cm open
☑ 10x18cm folded; 10x36cm open
☑ 10x21cm folded; 20x21cm open
☑ 10x6.5cm
☑ 11 1/2 x 9 1/2 in. fold. ; ill.
☑ 11 1/2 x 9 in.
☑ 11 1/2 x 9 in. fold. ; ill.
☑ 11 x 7 in.
☑ 11 x 7 in. ; ill.
☑ 11 x 8 1/2 in. ; ill.

☑ Auto Apply

6) Logical Mistakes

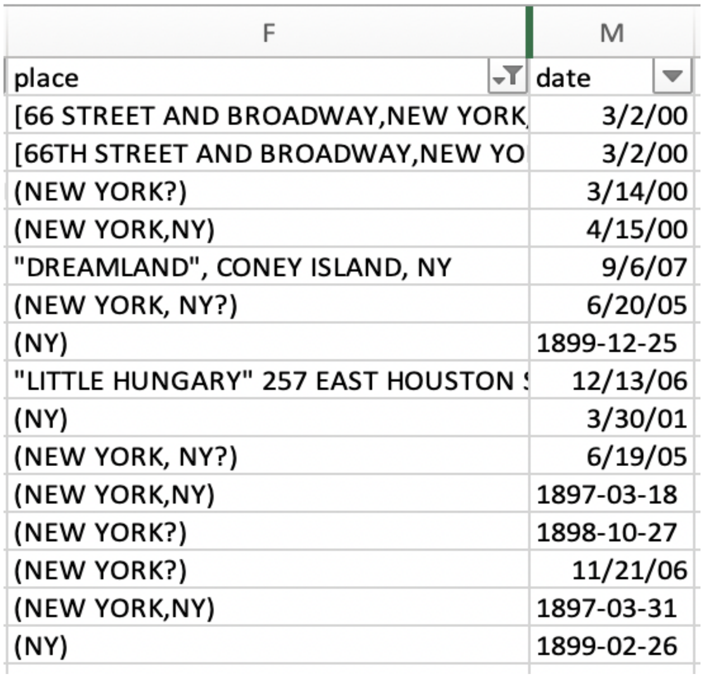
Some data are obviously wrong in logic. For example, “0” for price, “1” for first_appeared year, “2928” for first_appeared year or last_appeared year, “0001-01-01” for the date, etc.



	B	C	D	E	F	G	H	I	J
	name	description	menus_appeared	times_appeared	first_appeared	last_appeared	lowest_price	highest_price	
1	Consomme printaniere royal		8	8	1897	1927	0.2	0.4	
2	Chicken gumbo		111	117	1895	1960	0.1	0.8	
3	Tomato aux croutons		13	13	1893	1917	0.25	0.4	
4	Onion au gratin		41	41	1900	1971	0.25	1	
5	St. Emilion		66	68	1881	1981	0	18	
7	Radishes		3262	3346	1854	2928	0	25	
8	Chicken soup with rice		48	49	1897	1961	0.1	0.6	
9	Clam broth (cup)		14	16	1899	1962	0.15	0.4	
10	Cream of new asparagus, croutons		2	2	1900	1900	0	0	
11	Clear green turtle		157	157	1893	1937	0.25	60	
12	Striped bass saute, meuniere		2	2	1900	1900	0	0	
13	Anchovies		453	484	1858	1987	0	30	
14	Fresh lobsters in every style		4	4	1899	1900	0	0	
15	Celery		4246	4690	1	2928	0	50	
16	Pim-olas		145	148	1897	1918	0.15	35	
17	Caviar		505	534	1880	1987	0	75	
18	Sardines		1425	1484	1856	2928	0	50	
19	India chutney		16	16	1865	1901	0.1	0.2	
20	Pickles		453	472	1852	1987	0	10	
21	English walnuts		83	86	1851	1948	0.1	0.3	
22	Pate de foies-gras		9	9	1898	1901	1	1	
23	Pomard		11	11	1880	1950	0.75	5	
24	Brook trout, mountain style		2	2	1900	1900	0	0	
25	Whitebait, sauce tartare		4	4	1900	1901	0.3	0.75	
26	Clams		170	180	1881	1970	0.1	0.9	
27	Oysters		289	292	1862	1963	0	35	
28	Claremont planked shad		2	2	1899	1900	1.5	2.5	
29	G. H. Mumm & Co's Extra Dry		14	14	1895	1914	2	4	
30	Cerealine with Milk		1	1	1901	1901	0	0	
31	Sliced Bananas		220	238	1900	1987	0	15	
32	Wheat Vitos		3	3	1900	1900	0	0	
33	Sliced Tomatoes		1195	1256	1873	2928	0	25	
34	Russian Caviare on Toast		3	3	1900	1900			
35	Smoked beef in cream		21	21	1896	1933	0.3	0.9	
38	Apple Sauce		721	829	1	1987	0	20	
39	Potage a la Victoria		5	5	1899	1901			

7) Inconsistent format

This issue can be seen throughout the dataset. For example, as is shown in the following screenshot for the table “Menu”, the date format is inconsistent, some use ISO date format (YYYY-MM-DD), while others use different date formats, such as MM/DD/YY, YYYY, etc. It is the same case with the address format. Some listed the street, city, and country, while others listed the restaurant name instead. Some use the abbreviations, while others use the full forms.



place	date
[66 STREET AND BROADWAY,NEW YORK,	3/2/00
[66TH STREET AND BROADWAY,NEW YO	3/2/00
(NEW YORK?)	3/14/00
(NEW YORK,NY)	4/15/00
"DREAMLAND", CONEY ISLAND, NY	9/6/07
(NEW YORK, NY?)	6/20/05
(NY)	1899-12-25
"LITTLE HUNGARY" 257 EAST HOUSTON S	12/13/06
(NY)	3/30/01
(NEW YORK, NY?)	6/19/05
(NEW YORK,NY)	1897-03-18
(NEW YORK?)	1898-10-27
(NEW YORK?)	11/21/06
(NEW YORK,NY)	1897-03-31
(NY)	1899-02-26

In general, the data quality of this dataset is messy. We listed some obvious data quality issues spotted at our first glance. Surely, there are many other formatting or data quality problems, which will be addressed in our next phase, based on our use case and data cleaning goals, to achieve the desired “fitness for use”.

4 Use Case and Data Cleaning Goals

1) U0: The use case for which the original dataset is clean enough, thus data cleaning is not necessary

The descriptive data created for each entity when they were cataloged is digitally searchable, including useful information like the name of the restaurant, its geographical location, the physical description of its menu, etc. Although the dataset is dirty, we can use it to run some simple queries. For example, we can sort the column "times_appeared" in the table "Dish" to know which dish appeared the greatest number of times on the menus. The answer is "Coffee", which appeared 8484 times in total. It makes sense to me. We can also search the "status" of a certain menu by using the "menu id" and "status" attributes in the table "Menu". These 2 columns seem to be clean and complete. Another possible question the uncleaned dataset can help to answer is which dishes first appeared in 1900. We can get a list of the 20 dishes that first appeared in 1900 by running a quick query on the table "Dish".

2) U2: The use case for which the dataset will never be clean enough or usable, thus data cleaning is not sufficient

Data cleaning cannot resolve missing data or mis-transcribed data problems. Thus, the dataset cannot be used to correctly answer certain queries if the related information is incomplete or inaccurate. For example, there is a lot of missing data in the column "first_appeared" of the table "Dish". Thus, it may be impossible to answer a query such as "when the dish Celery first appeared on the menu", regardless how much data cleaning efforts we put in. The large amount of missing or unmatched values in the "price" related columns in the table "Dish" and the table "MenuItem" may also make a rigid study on price per each dish impossible.

3) U1: The target use case that matches our data cleaning goals, for which the data cleaning is both necessary and sufficient

After data cleaning, the dataset can be used for many use cases. The one we are interested in is to find out which restaurant in NY has the most menus collected in this dataset and which 10 dishes most often appeared on its menus. Another U1 use case we considered is to study the most popular dishes at a certain period for a certain event (e.g., breakfast).

The goal of our data cleaning work is to “fit-for-purpose”. We will link the four tables into a merged one, and then clean it to the extent that it can match our use case. We will demo our use case, analyze and visualize the results to show the difference brought by the data cleaning.

5 Tentative Plan

Obviously, the original dataset is not clean enough to match the use case we specified. We devised the following plan to guide our data cleaning work in the next phase.

Phase 1: This is the phase we are currently working on. In this phase, we will have an initial assessment of the dataset we selected. We will describe the structure, schema and content of the dataset. Meanwhile, we will inspect the data quality of the dataset. Obvious data quality problems will be spotted and listed. Also, we will develop the target use case and the matching data cleaning goals. Finally, a tentative work plan will be developed. Each task will be assigned to specific team members and its timeline will be scheduled.

Phase 2: This is the core phase of our project. In this phase, we will clean the dataset with OpenRefine or Python. After the data cleaning, we will develop the relational schema and check the integrity constraints with SQL. Finally, we will use yesWorkflow to establish a workflow model to tie all steps together.

Phase 3: This is the use case demo and documentation phase. In the use case demo, we will answer the questions we specified. We will use tableau to support the data analysis and visualization. As for the documentation, we will compare the cleaned dataset with the original dataset. Data quality changes will be listed. Answers to different queries will be compared. Lastly, we will summarize our challenges, findings, lessons learned and possible next steps.

Phase 4: This is the final report writing and project close phase. In this phase, we will document all our work, consolidate our report, and organize our deliveries. We will close the project and submit the report before the due date.

We use the project management tool - Gantt Chart to schedule and track our project processes. The project is broken down by phases and tasks. Each task is assigned to a team member. The tasks, responsible persons, start and end date, as well as the percent completed, are shown in the following chart.

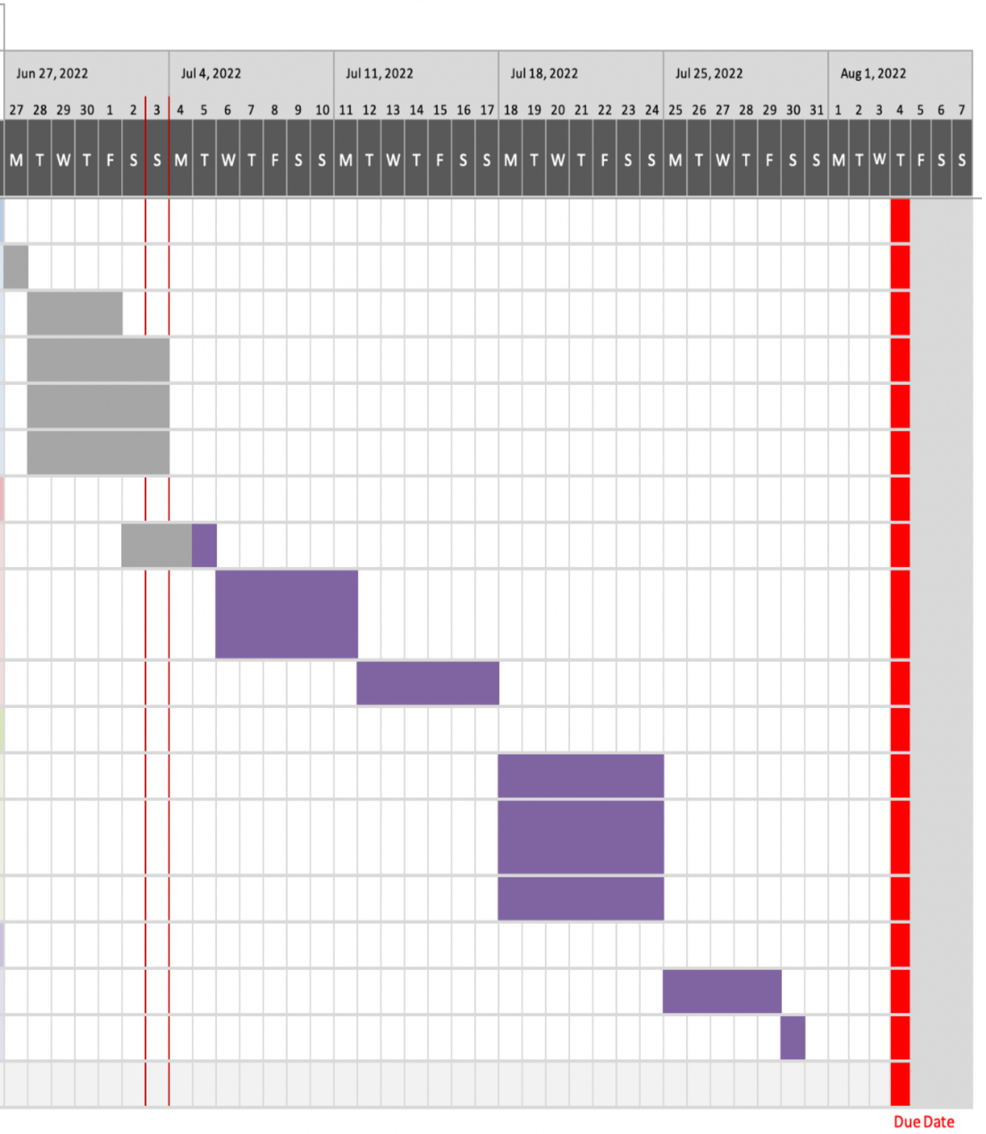
CS513: Theory & Practice of Data Cleaning

Final Project

Project Start: Mon, 6/27/2022

Display Week: 1

TASK	ASSIGNED TO	PROGRESS	START	END
Phase 1 Initial Assessment				
Select the dataset	All	100%	6/27/22	6/27/22
Describe the dataset: structure, schema and content	Jay	100%	6/28/22	7/1/22
Identify obvious data quality issues	Andy	100%	6/28/22	7/3/22
Develop use case and data cleaning goals	Jane	100%	6/28/22	7/3/22
Develop work plan: timeline and assignment of tasks	Jane	100%	6/28/22	7/3/22
Phase 2 Data Cleaning				
Clean the dataset - OpenRefine or Python	Jay	80%	7/2/22	7/5/22
Develop relational schema & check integrity constraints - SQL	Andy	0%	7/6/22	7/11/22
Creat a workflow model- yesWorkflow	Jane	0%	7/12/22	7/17/22
Phase 3 Use Case Demo and Documentation				
Demo the use case - tableau	Jay	0%	7/18/22	7/24/22
List data quality changes and compare results to specific quaries	Jane	0%	7/18/22	7/24/22
Challenges, findings, lessons learned, possible next steps	Andy	0%	7/18/22	7/24/22
Phase 4 Report Writing / Project Close				
Write final report and organize deliveries	All	0%	7/25/22	7/29/22
Submit the final work	All	0%	7/30/22	7/30/22


Jun 27, 2022
Jul 4, 2022
Jul 11, 2022
Jul 18, 2022
Jul 25, 2022
Aug 1, 2022
M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S M T W T F S S
Due Date