CS 513: Theory and Practice of Data Cleaning

Data Cleaning Project: Phase-1

Summer 2022

By:Team 31

Candice Yan

Email: yan40@illinois.edu

Yilin Hou

Email: yilin17@illinois.edu

Fangsheng Yang

Email: fyang28@illinois.edu

1. Identify a Dataset

The dataset is the menu and dish data developed by the New York Public Library.

2. Define main use case

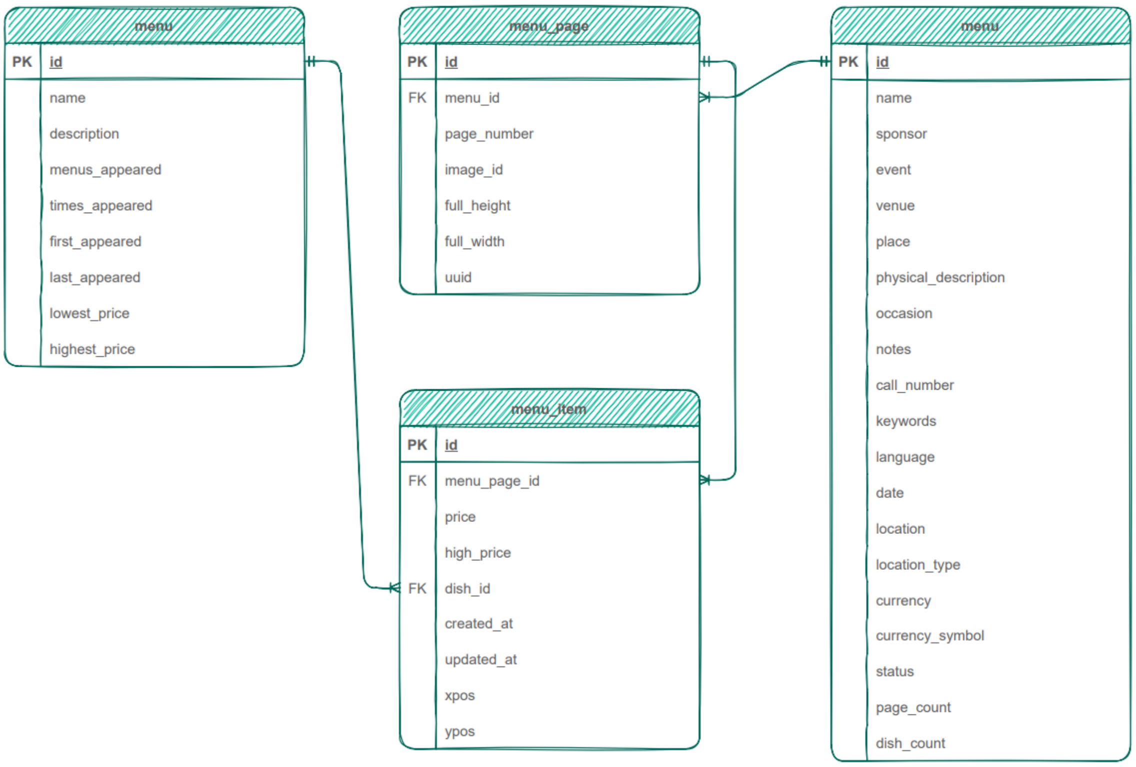
Use Case U0: No data cleaning needed when we are interested in Top 10 popular dishes on the menu. We can sort the menu_appeared column in a descending order and find the answers: Coffee, Tea, Celery, Olives, Radishes, Mashed potatoes, Milk, Boiled potatoes, Fruit, Chicken Salad.

Use Case U1: The menu dataset can be used to extract useful information about dishes. For instance, we can investigate popular dishes during different time periods to figure out changes of people's eating preference over time. This can be further broken down into more granular levels such as by venue or event.
Eating preferences are also highly related to culture variations and economic cycles - in good times, people usually eat outside more and thus we are expected to observe more dishes appearing on the menu, higher food prices and better dish qualities (e.g., more wine menus). In the downturn, less unnecessary consumption is made and we would see less dishes, lower prices and lower quality of food (e.g., basic protein and meat).

Use Case U2: the data cleaning is never enough when we would like to find which dish is the oldest in the full dish list because the first_appeared column can not be 100% clean. There is a count of 55287 for 0 and count of 205 for 1 on this column, so we never know what year these dishes first appeared on the menu. One solution is to find the oldest dishes on the menu excluding these 55,492 records.

3. Describe the Dataset

Diagram:



Schema:

```
--Table : dish
CREATE TABLE dish (
id int NOT NULL,
name varchar(255),
description varchar(255),
menus_appeared int,
times_appeared int,
first_appeared DATETIME,
last_appeared DATETIME,
lowest_price DOUBLE,
highest_price DOUBLE,
PRIMARY KEY (id),
);


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

Table : menu_page

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

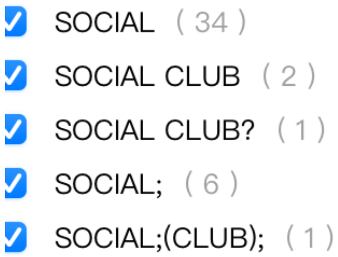
Table : menu

```
CREATE TABLE menu (
id int 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,
PRIMARY KEY (id)
);
```

4. Data Quality issues:

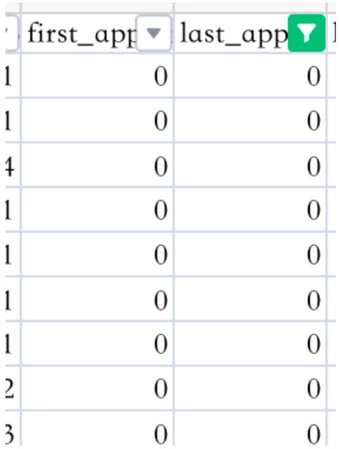
In order to get the dish table and menu table joint, we need to join through the menuitem and menupage table by menu_id, dish_id and menu_page_id (primary key). This will require integrity constraints not to be violated, thus a check on those id columns will be a must.

Venue and event are character fields, which contain ambiguous inputs and unconsolidated records.
For example, below columns could be further consolidated to social/social club.



- ☑ SOCIAL (34)
- ☑ SOCIAL CLUB (2)
- ☑ SOCIAL CLUB? (1)
- ☑ SOCIAL; (6)
- ☑ SOCIAL;(CLUB); (1)

Since we are analyzing dishes over different time periods, date will be a very important feature to mark the timeline. In the dish table, the Appeared date column has some missing values (~ 60k) as shown below, which is also worth to note. We could simply delete them as imputation could not be really appropriate.



	first_app	last_app
1	0	0
1	0	0
4	0	0
1	0	0
1	0	0
1	0	0
1	0	0
2	0	0
3	0	0

In the name column of the dish table, some are wrongly filled with years. That could be a transcribed error.
Date columns also suffer from large outliers - a year 2928 seems obviously odd and should be removed.

5. Initial Project Plan

- Describe the manu and dish dataset and associated use cases (done in Phase 1)

- Identify data quality problems that need to be solved for use cases (refer to 4. Data Quality Issues)
- Clean the data for date, name, venue, and event columns using OpenRefine
- Join tables and ensure integrity constraints are not violated using SQL
- Obtain results from SQL queries for defined use cases
- Check new dataset and compare it with the original dataset
- Document the change we have made from original dataset to cleaned dataset