Project Phase-I Report

Team36

Minh Nguyen – minhn2

Jiaqing Mo - jiaqing7

Haitham Shahin - hshahin2

Part 1: Description of Dataset

Database Schema

The NYPL Menu dataset consists of four interconnected CSV files representing a relational database structure:

Detailed Schema

Menu Table (17,545 records)

- Primary Key: id (integer)

- Attributes:

    - name (string) - Name of the menu/event
    - sponsor (string) - Organization or establishment hosting
    - event (string) - Type of meal (BREAKFAST, DINNER, LUNCH, etc.)
    - venue (string) - Type of venue (COMMERCIAL, HOTEL, etc.)
    - place (string) - Geographic location
    - physical_description (string) - Physical characteristics of menu
    - occasion (string) - Special occasions (EASTER, etc.)
    - notes (text) - Additional descriptive information
    - call_number (string) - Library catalog reference
    - keywords (string) - Search keywords
    - language (string) - Menu language
    - date (date) - Menu date (YYYY-MM-DD format)
    - location (string) - Specific location/establishment name
    - location_type (string) - Type of location
    - currency (string) - Currency type (e.g., "Dollars")
    - currency_symbol (string) - Currency symbol (e.g., "$")
    - status (string) - Processing status ("complete", "under review")

- page_count (integer) - Number of pages in menu
- dish_count (integer) - Number of dishes on menu

MenuPage Table (66,937 records)

• Primary Key: id (integer)

• Foreign Key: menu_id → Menu.id

• Attributes:

- page_number (integer) - Page sequence number
- image_id (integer) - Digital image identifier
- full_height (integer) - Image height in pixels
- full_width (integer) - Image width in pixels
- uuid (string) - Unique identifier for digital asset

MenuItem Table (1,332,726 records)

• Primary Key: id (integer)

• Foreign Keys:

  - menu_page_id → MenuPage.id
  - dish_id → Dish.id

• Attributes:

- price (decimal) - Menu item price
- high_price (decimal) - Upper price range (for variable pricing)
- created_at (timestamp) - Record creation timestamp
- updated_at (timestamp) - Last modification timestamp
- xpos (decimal) - X-coordinate position on menu page (0-1 normalized)
- ypos (decimal) - Y-coordinate position on menu page (0-1 normalized)

Dish Table (423,397 records)

• Primary Key: id (integer)

• Attributes:

- name (string) - Dish name
- description (text) - Dish description
- menus_appeared (integer) - Number of different menus featuring this dish
- times_appeared (integer) - Total occurrences across all menus
- first_appeared (integer) - Year of first appearance
- last_appeared (integer) - Year of last appearance
- lowest_price (decimal) - Minimum price recorded
- highest_price (decimal) - Maximum price recorded

Description

The New York Public Library (NYPL) Menu Collection represents a comprehensive digitization project of historical restaurant menus spanning from the mid-19th century to the early 21st century. The collection spans across 3 centuries, from 1800s to 2000s (it is worth noting there appears to be data error since earliest record of menu is year 01 and latest is later than current year of 2025), with the majority of menus dating from 1880-1980. This dataset includes a wide variety of venue types (commercial restaurants, hotels, etc.), meal types (breakfast, lunch, etc.) and physical metadata (detailed descriptions of menu design, materials, and artistic elements).

Part 2: Develop three use cases

Task 2a: Target Main Use Case (U1)

Scenario: How have dish prices changed with time and how do they differ across menu locations.

One hypothetical use case for the dataset is wanting to understand how the same dish is priced different both with variance in time and variance in menu location. This use case requires necessary data cleansing to first standardize the set of dishes such that syntax issues or differences in language used for a given dish that represent the same semantic dish are matched. Next, the data needs to be validated to ensure that year and price are valid values in the dataset. Additionally, the menu locations and venues need to be validated as real locations on the map. It is also necessary to ensure that dishes can be mapped back to their origin menus since this link is how a given dish can be compared across locations.

Query Samples (In Natural Language):

1) Select avg(price) for dishes grouped by decades for a given country -> This helps to show how in a particular location dish prices changed over time

2) Select top(prices) for each dish in a given decade, country -> helps to show how prices changed for a dish in a given country over each decade

3) Select avg(prices) for each country across decades -> show how prices changed across dishes in each country over time

Task 2b: "Zero Data Cleaning" Use Case (U0)

Scenario: A library science researcher wants to analyze the digitization quality and completeness of the NYPL menu collection to assess the success of the crowdsourcing project and identify patterns in volunteer contributions.

The analysis requirements are to (1) analyze completion status using the status field ("complete" vs "under review") and (2) compare digitization quality metrics using page_count and dish_count fields.

Zero data cleaning is sufficient because (1) the status field provides direct categorical data ("complete", "under review") that requires no standardization for analysis and (2) the page_count and dish_count fields contain clean numerical data suitable for statistical analysis without modification. In Menu.csv, there are no rows with empty values in status, page_count and dish_count columns.

Task 2c: "Never Enough" Use Case (U2)

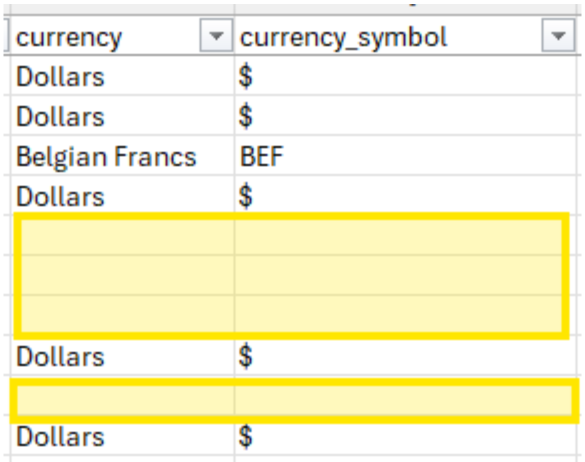
Scenario: A public health researcher wants to conduct a comprehensive nutritional analysis of American dining patterns over the past 150 years to model the evolution of dietary habits and nutritional content. The goal is to correlate historical dietary patterns with health outcomes and inform modern public health policy.

The analysis requirements are to (1) calculate nutritional content (calories, macronutrients, vitamins, minerals) for each dish and (2) standardize ingredient lists and cooking methods across time periods

Data cleaning is never sufficient because (1) The dataset contains dish names and brief descriptions but lacks the detailed ingredient lists, portion sizes, and preparation methods necessary for nutritional analysis. For example, "Chicken gumbo" appears 117 times across different establishments and time periods, but without standardized recipes, nutritional content cannot be determined. (2) Menu descriptions like "Tomato aux croutons" or "Cream of new asparagus, croutons" provide insufficient detail about ingredients, quantities, or preparation methods needed for nutritional analysis. In short, there are fundamental data gap because there is no ingredient quantities or preparation methods.

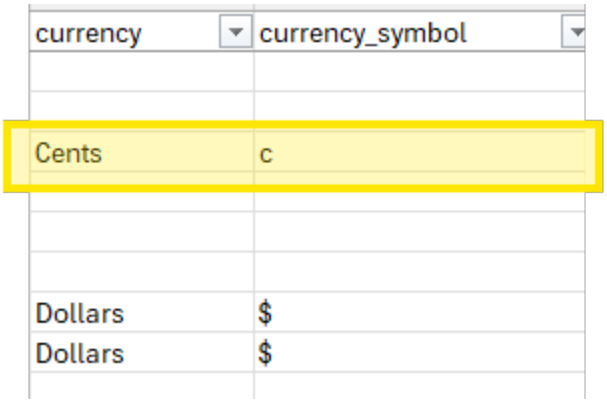
Part 3: List obvious data quality problems

1. In the Menu table, there are records that are missing currency and/or currency symbol values, making it impossible for price comparison and analysis of these records.



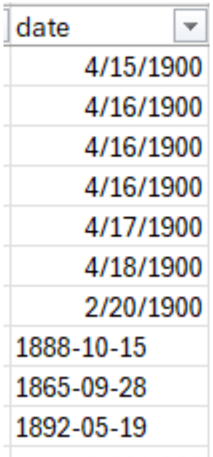
currency	currency_symbol
Dollars	$
Dollars	$
Belgian Francs	BEF
Dollars	$
Dollars	$
Dollars	$

2. In the Menu table, there are records that have invalid currency and/or currency symbol values, like "Cents", making it impossible for price comparison and analysis of these records.



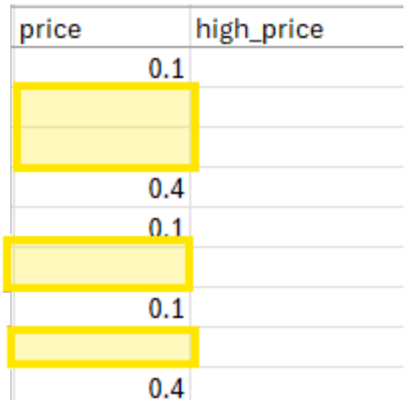
currency	currency_symbol
Cents	c
Dollars	$
Dollars	$

3. In the Menu table's date column, there are different formats of dates, so we cannot compare the dates for all the records.



date
4/15/1900
4/16/1900
4/16/1900
4/16/1900
4/17/1900
4/18/1900
2/20/1900
1888-10-15
1865-09-28
1892-05-19

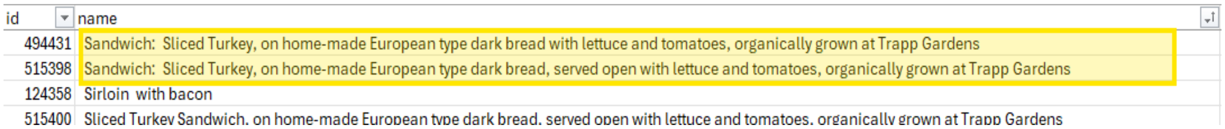
4. In the MenuItem table, there are records that are missing price values, making it impossible for price comparison and analysis of these records.



price	high_price
0.1	
0.4	
0.1	
0.1	
0.4	

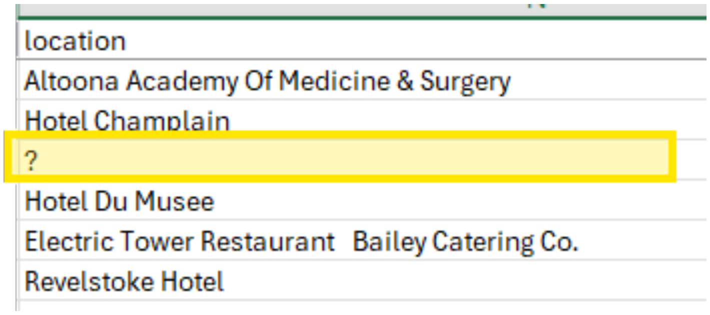
5. Integrity Constraint check is necessary to make sure all menu items have a corresponding dish and menu record.

6. To properly do dish price comparisons, it will be necessary to standardize/aggregate certain dishes that represent the same dish but may have different spellings or references.



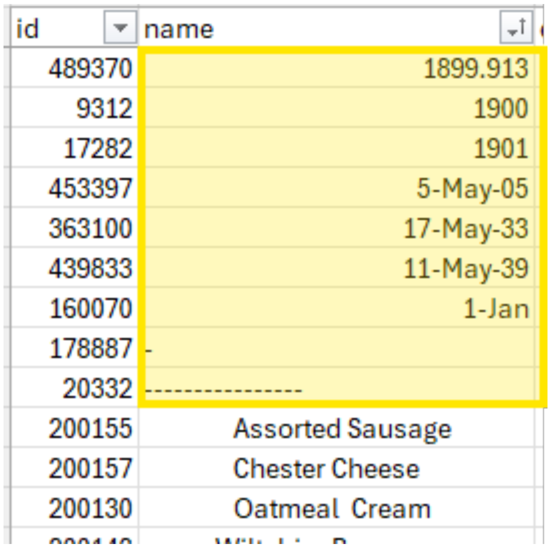
id	name
494431	Sandwich: Sliced Turkey, on home-made European type dark bread with lettuce and tomatoes, organically grown at Trapp Gardens
515398	Sandwich: Sliced Turkey, on home-made European type dark bread, served open with lettuce and tomatoes, organically grown at Trapp Gardens
124358	Sirloin with bacon
515400	Sliced Turkey Sandwich, on home-made European type dark bread, served open with lettuce and tomatoes, organically grown at Trapp Gardens

7. Menu venues/locations will need to be checked for accuracy on a geographic map or against a real/historical address. This will allow for comparison across locations.



location
Altoona Academy Of Medicine & Surgery
Hotel Champlain
?
Hotel Du Musee
Electric Tower Restaurant Bailey Catering Co.
Revelstoke Hotel

8. There are some invalid dish names in the Dish table. If we can’t determine what the actual names of the dishes are, then they can’t be used in the analysis.



id	name
489370	1899.913
9312	1900
17282	1901
453397	5-May-05
363100	17-May-33
439833	11-May-39
160070	1-Jan
178887	-
20332	----------------
200155	Assorted Sausage
200157	Chester Cheese
200130	Oatmeal Cream

Part 4: Initial Data Cleansing Plan for Phase 2

Step 1) Review the Use Case Description and Dataset Description.
The goal of this step to validate that the dataset, once cleaned, can support the use case. This includes reviewing the dataset descriptions and the questions/queries that are intended to be addressed in the use case.

- Assigned To: Haitham Shahin

- Timeline: 07/21

Step 2) Profile the Dataset to Identify Data Quality Problems with SQL.
Through the use of SQL, the team will show data quality problems that exist in the uncleaned Dataset that need to be addressed to answer the use case.

- Assigned To: Haitham Shahin

- Timeline: 07/21

Step 3) Perform the Data Cleansing Steps.
This is executing the data cleaning steps that have been identified as necessary and sufficient to address the use case. Tools include SQL, Python, OpenRefine, and YesWorkflow to execute the full data cleaning workflow. The full team will participate in this effort. As Steps 1 and 2 are completed, the team will identify the critical steps for the cleaning and distribute the steps of the workflow across the team members.

- Timeline: 07/28

Step 4) Data Quality Checking to Show the Cleansed Dataset addresses the necessary and sufficient steps to address the use case.
The team will show how the cleansed dataset now can be sufficiently used to answer the use case.

- Assigned To: Jiaqing Mo

- Timeline: 07/30

Step 5) Document and Quanity the Changes Made to the Dataset
This step is to demonstrate and report on the cleansing applied to the dataset.

- Assigned To: Minh Nguyen

- Timeline: 07/30