CS513 Project Phase-I Report

Team69 (Team Unicorn)
Zhao Li (zhaol4@illinois.edu)
Kaihan Shen (kaihans2@illinois.edu)
Chen Yuan (cheny9@illinois.edu)

1. Dataset Chosen (5 points)

In this project, we will use the NYPL-menus dataset (https://uofi.app.box.com/s/whvfh9jio38ck0m9gz58s31srx8iwg4i/folder/159094620210).

2. Description of Dataset

The dataset we will use is New York Library's menu collection. It contains more than 45,000 menus dating from the 1840s to present. The dataset contains menus manually transcribed by volunteers such that users such as historians, chefs and food enthusiasts can access menu information easily.

The data we downloaded from the website contains four csv files:

Dish.csv

Dish(id, name, description, menus_appeared, times_appeared, first_appeared, last_appeared, lowest_price, highest_price)

This table contains information about a dish. The dish id is the primary key of the table, and it contains information such as name of the dish, description of the dish, how many menus/times it appeared, first/last time appeared and lowest/highest price.

Menu.csv

Menu(id, name, sponsor, event, venue, place, physical_description, occasion, notes, call_number, keywords, language, date, location, location_type, currency, currency_symbol, status, page_count, dish_count)

This table contains features of the menu. The menu id is the primary key of the table. This table contains information such as name/sponsor of the menu, geographical location, currency and page/dish count.

MenuItem.csv

MenuItem(id, menu_page_id, price, high_price, dish_id, created_at, updated_at, xpos, ypos)

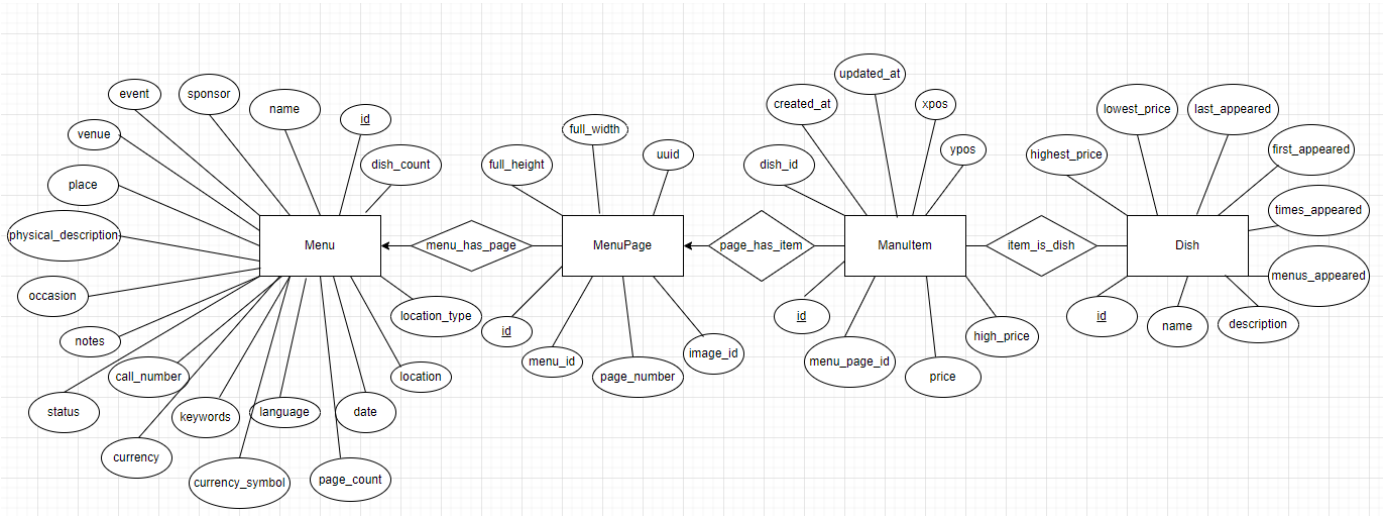
This table contains all the items in each menu. It also links between a dish and a menu page. The id is the primary key of the table. Menu_page_id is a foreign key to the id

column in MenuPage table and dish_id is a foreign key to the id column in Dish table.The table contains information such as price and location of the item.

MenuPage.csv
MenuPage(id, menu_id, page_number, image_id, full_height, full_width, uuid)

This table contains information about each page in a menu. The menu page id is the primary key of the table. The menu id column is a foreign key to the id column in Menu table.The table contains information such as size of the page, page number and image/uuid.

An ER diagram that illustrates the relationship between these tables is:



3. Use Cases

a. “Zero cleaning” use case U0:

Find the menu with the most menu pages. (MenuPage is a rather clean dataset and each data record has a corresponding page number, we can easily find the specific menu with the most pages without cleaning)

b. “Main” use case U1:

Predicting the popularity of a dish based on its historical appearance, event and price range. (Data cleaning and preprocessing is significant for this use case considering any invalid data might have impact on the validity and accuracy of this model)

c. “Never enough” use case U2 :

Find the mostly supported language among all menus. (Since the “language” column is empty, we won’t be able to find the language with any data cleaning).

4. Data Quality Problems

a. List obvious data quality problems with evidence (examples and/or screenshots) (20 points)

1. In the menu dataset, the columns notes, language, call_number, currency, currency_symbol missing data for most rows. And keywords missing data for all rows.

```
: menu.isna().sum() / menu.shape[0]
```

```
: id                    0.000000
  name                  0.817783
  sponsor               0.088971
  event                 0.535252
  venue                 0.537247
  place                 0.537019
  physical_description  0.158564
  occasion              0.783927
  notes                 0.395098
  call_number           0.089028
  keywords              1.000000
  language              1.000000
  date                  0.033400
  location              0.000000
  location_type         1.000000
  currency              0.632032
  currency_symbol       0.632032
  status                0.000000
  page_count            0.000000
  dish_count            0.000000
  dtype: float64
```

Also for the menuItem dataset it has missing values in high_price for most rows.

```
menuitem.isna().sum() / menuitem.shape[0]
```

```
id              0.000000
menu_page_id    0.000000
price           0.334589
high_price      0.931040
dish_id         0.000181
created_at      0.000000
updated_at      0.000000
xpos            0.000000
ypos            0.000000
dtype: float64
```

2. In the dish dataset, the 'last_appeared' column for some rows have incorrect year (e.g '2928)', which is a future year

	id	description	menus_appeared	times_appeared	first_appeared	last_appeared	lowest_price	highest_price
count	423397.000000	0.0	423397.000000	423397.000000	423397.000000	423397.000000	394297.000000	394297.000000
mean	264456.594900	NaN	3.060489	3.146794	1675.514555	1679.299738	0.965265	1.603875
std	150489.070889	NaN	27.818178	29.962122	651.321461	651.934580	6.714564	12.696274
min	1.000000	NaN	0.000000	-6.000000	0.000000	0.000000	0.000000	0.000000
25%	132374.000000	NaN	1.000000	1.000000	1900.000000	1900.000000	0.000000	0.000000
50%	269636.000000	NaN	1.000000	1.000000	1914.000000	1917.000000	0.000000	0.000000
75%	397135.000000	NaN	1.000000	1.000000	1949.000000	1955.000000	0.400000	0.600000
max	515677.000000	NaN	7740.000000	8484.000000	2928.000000	2928.000000	1035.000000	3050.000000

3. The 'lowest_price' and 'highest_price' in dish dataset do not specify the currency used.

b. Explain why / how data cleaning is necessary to support the main use case U1 (10 points)

Data cleaning is necessary because data cleaning ensures that only valid and accurate data is used in the model. Incorrect data can lead to false predictions and misleading insights. For example, a dish that mistakenly appears to have a future date in the 'last_appeared' column (e.g 'dish' dataset for 'Radishes'(id=7)) would lead to wrong calculations and predictions.

Also, missing data can distort the results of the analysis. For example, missing high_price (as seen in the 'menuItem' dataset) can affect the analysis of the dish's price range.

As discussed above, data cleaning helps reduce errors and avoid further potential issues. It reduces the risk of drawing incorrect conclusions from the data.

5. Initial Plan for Phase-II (10 points)

■ S1: Review (and update if necessary) your use case description and dataset description

This step has been completed in Phase I Report. We have described the dataset, identified potential data quality issues and provided potential use cases.

Team members: All team members.

Timeline: Completed by July 8th.

■ S2: Profile D to identify DQ problems: How do you plan to do it? What tools are you going to use?

Our use cases will use both numerical and text information from the data. Therefore, for numerical values, our plan is to check missing values and outliers, and identify any values that are suspicious. For example, some dishes with 0 price or an extremely high price certainly don't make any sense. For text data such as events, we will do simple tests, such as run value counts or do clustering, to see if there’re misspellings or leading/trailing spaces.

The tools we plan to use are Python and OpenRefine.

Team members: Chen Yuan, Kaihan Shen

Timeline: Expected to be completed by July 16th.

■ S3: Perform DC “proper”: How are you going to do it? What tools do you plan to use? Who does what?

For numerical variables, we plan to replace/remove suspicious values such as missing values and outliers. For text information, we will remove punctuations and leading/trailing spaces and use clustering methods to identify and combine the same events.

For numerical variables, we plan to use Python(Pandas, Numpy, etc.), OpenRefine and SQLite, and for text information, our plan is to use OpenRefine and SQLit.

Team members:Kaihan Shen, Zhao Li
Timeline: Expected to be completed by July 23rd.

- S4: Data quality checking: is D' really "cleaner" than D?

  We will write a Python script to read all the processed data to make sure no above quality issues exist anymore, for example, we will have scripts to check missing values, duplicates, foreign key constraints, etc. We will also have demos to present the manual validation process.

  Team members: Chen Yuan, Zhao Li
  Timeline: Expected to be completed by July 30th.

- S5: Document and quantify change

  We will write a very detailed report, which includes all the data cleaning steps we have done on the data, the comparison between old dataset and new dataset and our conclusions and findings on the dataset.

  Team members: Kaihan Shen, Zhao Li, Chen Yuan
  Timeline: Expected to be completed by July 30th.