Team52 - Arnab, Kavithaa, Sudipto

Data Cleaning Project Phase I:

1) **Identify a Dataset:** NYPL dataset for Menus (Menu.csv)

2) **Develop a Use Case :**

   Target (main) use case U1

   Cleaning the dataset is necessary for supporting the following data analysis use cases.

   i) Events can be filtered correctly and efficiently to get appropriate data for the various events at different restaurants in the dataset.However, in order to see whether the type of the event is Breakfast or Dinner or Lunch we need to clean the data.
   ii) To group venues in the menu dataset , cleaning is needed as different spellings for the same venue type exist. Example :Social has different spellings life SOC,SOC?, etc.
   iii) To sort or group the menus by occasion, cleaning is needed.Example: ANNIV, ANNIVERSARY?etc
   iv) To sort or group the menus by place, cleaning is needed. Example Chicago has many different spellings.

   b)**U0-zero data cleaning**

   We can get the total number of completed orders (Status = 'completed') and who was the sponsor if any for those orders without doing any data cleaning.
   Page_count and dish_count columns will not require any cleaning.
   Sample Query for Page_count:
   Select * from menus where Page_count>3;

   c)**U2 -never (good) enough**

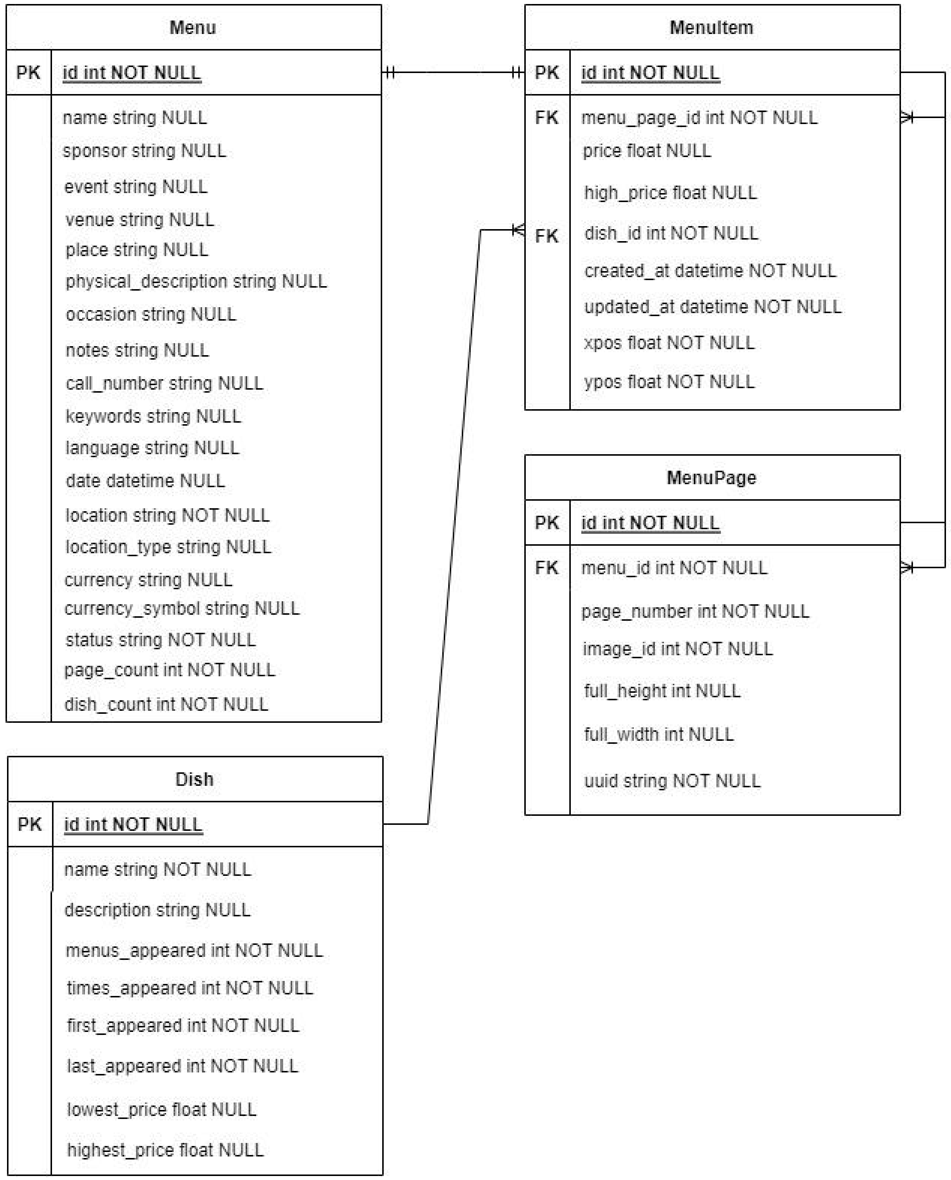
   Name,Call_number, Date,Sponsor columns are missing few rows ,so even cleaning the data will not provide correct answers for the queries

3) **Describe the Dataset:**

   The Menu Dataset is created by the New York Public Library by collecting various menus from 1840 to present date. The dataset has 17,400 menus from across the world. Each table has multiple attributes.

   The dataset highlights the details of the menus, food catering service that took place at different events at different venues across the world. It shows the Sponsor for the event, the event type, the venue details, the menu type for that event with notes, occasion for the event, date when it was held, number of dishes served and the status of the event. It also has the supporting tables showing the

details for the Menu, Food/Dish served and the page details in each Menu. Below is the ER diagram and a high level description of each table in the entire dataset.
ER Diagram:



Explanation of tables in dataset:

MenuItem: (Highlights the individual Menu items present in a Menu with the dish served as part of the menu and its price in the menu)
Id - Primary key for MenuItem table
menu_page_id - Id for the menu page which has the Menu Item
price - Price of the Menu Item
high_price - Highest price amongst the menu item
dish_id - Dish Id for the dishes in the Menu
created_at - Date when Menu was created

updated_at - Last date when Menu was updated
xpos - X Position in menu
ypos - Y Position Menu

MenuPage: (Highlights the details of each page in a Menu book depicting the individual menu items present in that page)
id - Unique id to identify the Menu Page within a Menu
Menu_id - The menu id where the page belongs
page_number - Page number within the menu
image_id - image id for the menu
Full_height - height of the menu
full_width - width of the menu
uuid - UID for the menu page

Dish: (Highlights individual dishes that goes into different menus along with statistical data of their occurrences in those menus)
id - Unique id for the dish
name - Name for the dish
description - Description of the dish
menus_appeared - the Menu Ids in which the dish has occurred
times_appeared - no of times it appeared in any menu
first_appeared - date when it first appeared in a menu
last_appeared - date when it last appeared in a menu
lowest_price - lowest price of the dish in any of the menu
highest_price - highest price of the dish in any of the menu

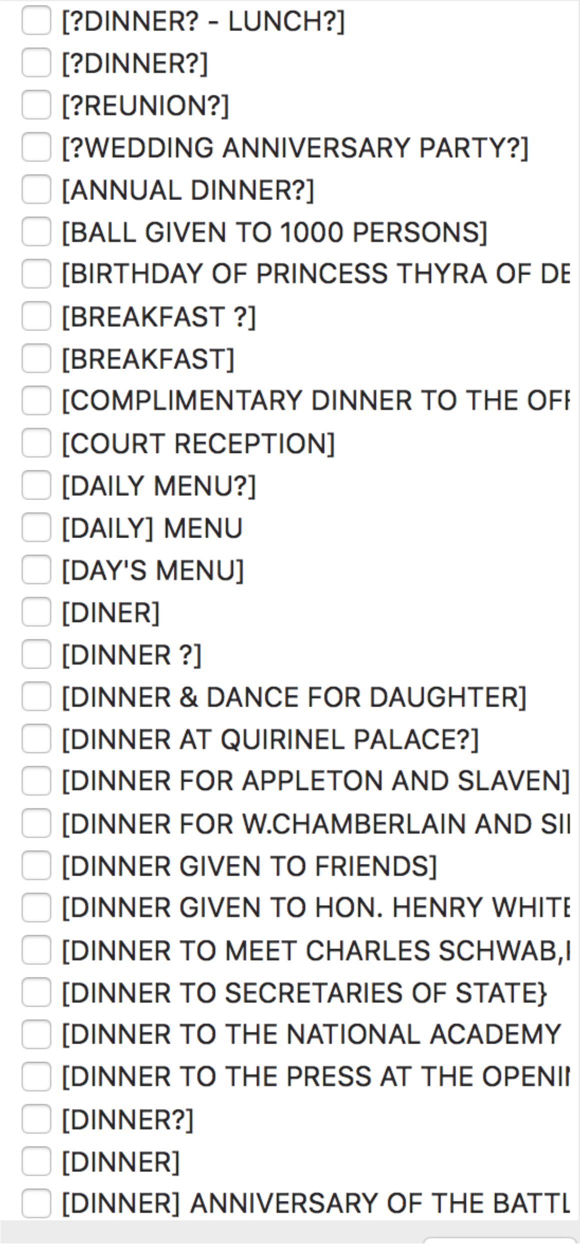
Menu: (Depicts the whole Menu itself and shows the different events/locations at different times when this menu was used)
id - Primary key for Menu table
name - Name of the restaurant
sponsor - name of the restaurant/hotel
event - the occasion
venue - the type of venue
place - street address, city, state
physical_description - physical description of menu_id
occasion - occasion of the menu (wedding, birthday etc)
notes - curator's note about the menu
call_number - the number to dial in
keywords - keywords to identify the menu
language - language used in menu
date - the menu date
location - event location (usually restaurant/hotel)
location_type - type of the location
currency - currency listed on the menu

currency_symbol - currency symbol as it appears on the menu
status - status of the data curation progress (complete or under review)
page_count - the number of pages in the menu
dish_count - the total dishes in the menu

4) List the Obvious Data Quality Problem

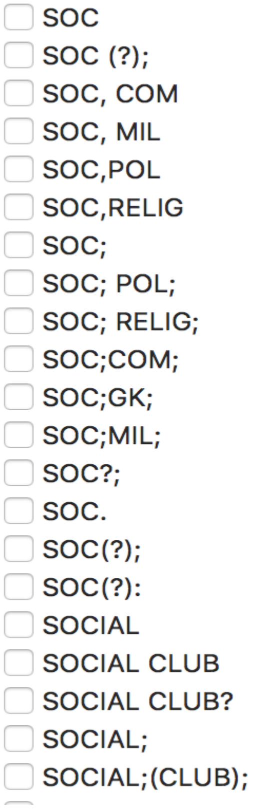
The dataset has quite a bit of data quality issues which are as below:

i) Event column describing the event type has redundant names like DINNER, [DINNER]. It also has spelling errors for the same Event column like DINNE in place of DINNER



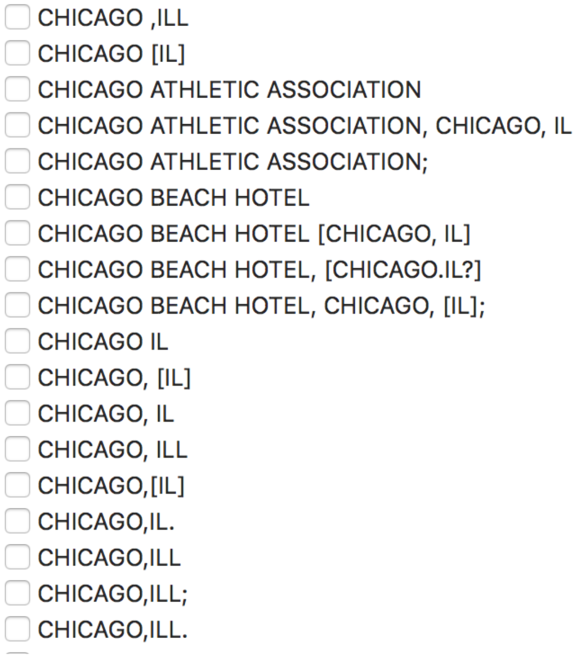
□ [?DINNER? - LUNCH?]
□ [?DINNER?]
□ [?REUNION?]
□ [?WEDDING ANNIVERSARY PARTY?]
□ [ANNUAL DINNER?]
□ [BALL GIVEN TO 1000 PERSONS]
□ [BIRTHDAY OF PRINCESS THYRA OF DE
□ [BREAKFAST ?]
□ [BREAKFAST]
□ [COMPLIMENTARY DINNER TO THE OFF
□ [COURT RECEPTION]
□ [DAILY MENU?]
□ [DAILY] MENU
□ [DAY'S MENU]
□ [DINER]
□ [DINNER ?]
□ [DINNER & DANCE FOR DAUGHTER]
□ [DINNER AT QUIRINEL PALACE?]
□ [DINNER FOR APPLETON AND SLAVEN]
□ [DINNER FOR W.CHAMBERLAIN AND SII
□ [DINNER GIVEN TO FRIENDS]
□ [DINNER GIVEN TO HON. HENRY WHITE
□ [DINNER TO MEET CHARLES SCHWAB,I
□ [DINNER TO SECRETARIES OF STATE}
□ [DINNER TO THE NATIONAL ACADEMY
□ [DINNER TO THE PRESS AT THE OPENII
□ [DINNER?]
□ [DINNER]
□ [DINNER] ANNIVERSARY OF THE BATTL

ii) The same is the case with venue, Social has many different spellings like SOC,SOC(?) etc



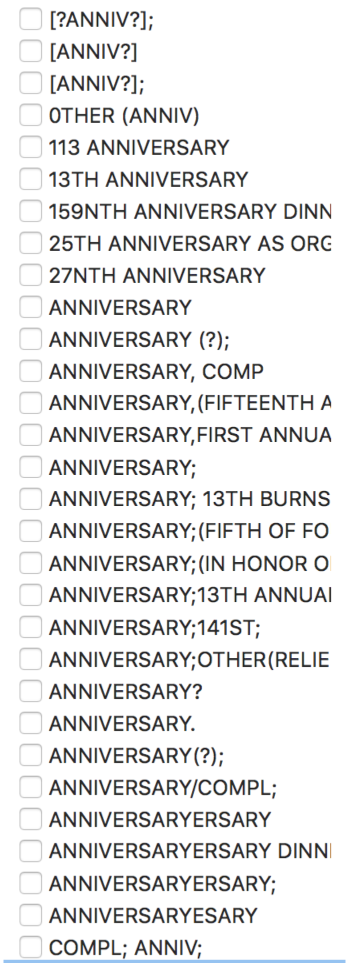
- □ SOC
- □ SOC (?);
- □ SOC, COM
- □ SOC, MIL
- □ SOC,POL
- □ SOC,RELIG
- □ SOC;
- □ SOC; POL;
- □ SOC; RELIG;
- □ SOC;COM;
- □ SOC;GK;
- □ SOC;MIL;
- □ SOC?;
- □ SOC.
- □ SOC(?);
- □ SOC(?):
- □ SOCIAL
- □ SOCIAL CLUB
- □ SOCIAL CLUB?
- □ SOCIAL;
- □ SOCIAL;(CLUB);

iii) The place columns as well where there are values like ‘“CHICAGO,[IL] ”,”CHICAGO,ILL”



- □ CHICAGO ,ILL
- □ CHICAGO [IL]
- □ CHICAGO ATHLETIC ASSOCIATION
- □ CHICAGO ATHLETIC ASSOCIATION, CHICAGO, IL
- □ CHICAGO ATHLETIC ASSOCIATION;
- □ CHICAGO BEACH HOTEL
- □ CHICAGO BEACH HOTEL [CHICAGO, IL]
- □ CHICAGO BEACH HOTEL, [CHICAGO.IL?]
- □ CHICAGO BEACH HOTEL, CHICAGO, [IL];
- □ CHICAGO IL
- □ CHICAGO, [IL]
- □ CHICAGO, IL
- □ CHICAGO, ILL
- □ CHICAGO,[IL]
- □ CHICAGO,IL.
- □ CHICAGO,ILL
- □ CHICAGO,ILL;
- □ CHICAGO,ILL.

iv) The Occasion columns as well where there are values like ‘“[ANNIV],[ANNIV?] ”



- □ [?ANNIV?];
- □ [ANNIV?]
- □ [ANNIV?];
- □ 0THER (ANNIV)
- □ 113 ANNIVERSARY
- □ 13TH ANNIVERSARY
- □ 159NTH ANNIVERSARY DINN
- □ 25TH ANNIVERSARY AS ORG
- □ 27NTH ANNIVERSARY
- □ ANNIVERSARY
- □ ANNIVERSARY (?);
- □ ANNIVERSARY, COMP
- □ ANNIVERSARY,(FIFTEENTH A
- □ ANNIVERSARY,FIRST ANNUA
- □ ANNIVERSARY;
- □ ANNIVERSARY; 13TH BURNS
- □ ANNIVERSARY;(FIFTH OF FO
- □ ANNIVERSARY;(IN HONOR O
- □ ANNIVERSARY;13TH ANNUAI
- □ ANNIVERSARY;141ST;
- □ ANNIVERSARY;OTHER(RELIE
- □ ANNIVERSARY?
- □ ANNIVERSARY.
- □ ANNIVERSARY(?);
- □ ANNIVERSARY/COMPL;
- □ ANNIVERSARYERSARY
- □ ANNIVERSARYERSARY DINNI
- □ ANNIVERSARYERSARY;
- □ ANNIVERSARYESARY
- □ COMPL; ANNIV;

5) Devise an Initial Plan -

We found there are about 17456 menu records in the datasheet, this datasheet was put together by the New York Times.
There are several use cases we are trying to solve here where the Menu table data is not appropriate -
Event type such as Breakfast, Dinner, Lunch, etc.
Group Venues in the Menu dataset.
Occasion.
Place or city names.

In order to use these data properly, we will have to run the clean up job on these columns. Primarily we will be using the tool called OpenRefine and we will take help of Regular Expressions(Regex) as well in order to fix this data.

Once the data is fixed, we will be using SQLite3 db to store that data. And we will use Python jobs to automate the whole process.

Once the data is cleaned we will be able to get the menu items for each event type. For example, we can query menu items for Breakfast for Social venues in a particular city like Chicago.

Since the data is cleaned now, the Breakfast, Social and Chicago values are consistent and standard across all records.

We will save the OpenRefine logs and we will write queries to show how many rows has changed as a result of this.

User stories -

- Use the Menu table to extract the relevant columns and identify the problematic values from those columns - Kavithaa
- Identify the Regex for cleanup - Arnab
- Use Openrefine and clean the data - Arnab, Kavithaa, Sudipto
- Verify manually that the data is cleaned - Kavithaa, Sudipto
- Write a Python script to load the cleaned data into a database - Arnab
- Documentation - Sudipto