**Data Clorox  
CS513 Data Cleansing**

Final Project Report

July 5, 2020

Asad Bin Imtiaz, Muhammad Rafay

# Overview and initial assessment of the dataset

In the following subsection, the structure and content on the dataset is inspected before starting with the data wrangling and provenance workflow, to get familiarity with data schema and a feel for apparent data quality issues present in the data. There may be more issues in data that would be discussed in subsequent chapters with corresponding tasks. The initial assessment was performed to get an understanding of the data quality in general and to identify methodology and tools for subsequent tasks.

## Data Files

The entire dataset consists of four character-delimited files described below:

1. ***Dish.csv***

This file contains dish names listed on the menu along with their respective pricing and chronology information. Each record represents a specific dish offered by a business and listed on the menu. Each dish has an identifier that uniquely identifies it and is referenced as a foreign key on other entities.

1. ***MenuItem.csv***

This file contains menu items that link a menu page entity with dish entities as foreign references. Each record is identified by a unique identifier and carries other information such as associated dish price and x/y position of the image of the menu page.

1. ***MenuPage.csv***

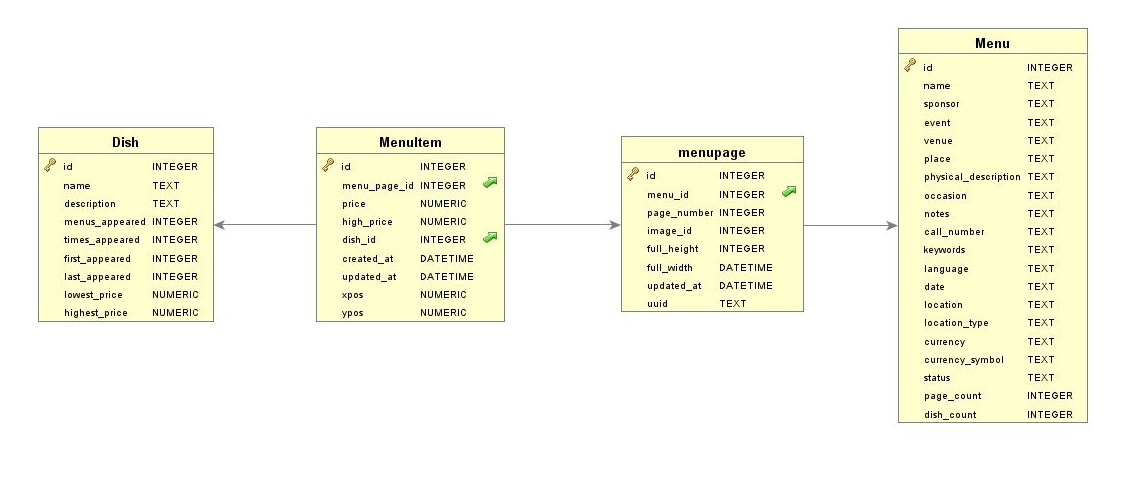
This file contrains menu page records. Each item is identified by a unique identifier and links a menu item with a menu. Additional information such as page photo image number and page dimensions also appear here. Every record keeps references of the menu item identifier and menu identifier to link these entities together.

1. ***Menu.csv***

This file contains all individual menus, each associated with a unique id. Each menu has an identifier that uniquely identifies it and is referenced as a foreign key on other entities. Associated data includes the occasion, venue and event information and chronological information such as created and updated dates and times. Other important fields present in this file include the location where the menu is offered, the associated currency in use for the menu items, the language for the menu and the status, among others.

## Data Structure & Schema

The raw data were imported in an SQL-Lite instance and visualized using the DB-Visualizer tool. The ER diagram generated from DB-Visualizer is shown in the figure below:



The cardinalities of objects with respect to one another are found to be (only list for entities among where a direct link is possible):

|  |  |  |
| --- | --- | --- |
| From Entity | To Entity | Cardinality of relation |
| Dish | MenuItem | 1:N |
| MenuItem | Dish | 1:1 |
| ManuItem | MenuPage | 1:1 |
| MenuPage | MenuItem | 1:N |
| MenuPage | Menu | 1:1 |
| Menu | MenuPage | 1:N |

Given the cardinalities above and the Initial Quality Assessment of the dataset in section 2.4, we have assessed that some rows are duplicate in the data and they need to be merged according to certain criteria.

e.g in Dish.csv the following dish has a duplicate due to difference in cases

Dish.id = 1 , Dish.name = 'Consomme printaniere royal'

Dish.id = 397198, Dish.name = 'Consomme Printaniere Royal'

Although the cardinality from MenuItem to Dish of 1:1 is maintained due to the unique Dish.id but just due to difference in cases for dish names, there should not exist multiple records representing the same dish. These problems can only be analyzed and mitigated after initial cleansing.

## Data Types

The diagram shows entities and links for data objects present in the data set. Most of the raw data was imported as strings of characters. However, the initial assessment showed the following data types for the fields:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **Dish** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| Id | Integer |  | (10)9 | PK | N |
| Name | String | 1387 | X(1387) unicode |  | N |
| Description | String | 0 | X(1) |  | Y |
| Menus\_appeared | Integer |  | -(10)9 |  | N |
| Times\_Appeared | Integer |  | -(10)9 |  | N |
| First\_Appeared | Integer |  | (4)9 |  | N |
| Last\_Appeared | Integer |  | (4)9 |  | N |
| Lowest\_Price | Numeric | 2 | --------.99 |  | Y |
| Highest\_Price | Numeric | 2 | --------.99 |  | Y |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **MenuItem** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| Id | Integer |  | (10)9 | PK | N |
| Menu\_Page\_Id | Integer |  | (10)9 | FK | N |
| Price | Numeric | 2 | ----.99 |  | Y |
| High\_price | Numeric | 2 | ----.99 |  | Y |
| Dish\_id | Integer |  | (10)9 | FK | Y |
| Created\_at | Timestamp(0)  With zone |  | YYYY-MM-DD HH:MM:SS(0) Z |  | N |
| Updated\_at | Timestamp(0)  With zone |  | YYYY-MM-DD HH:MM:SS(0) Z |  | N |
| Xpos | Numeric | 6 | -.999999 |  | N |
| Ypos | Numeric | 6 | -.999999 |  | N |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **MenuPage** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| Id | Integer |  | (10)9 | PK | N |
| Menu\_Id | Integer |  | (10)9 | FK | N |
| Page\_Number | Integer |  | ----.99 |  | Y |
| Image\_Id | String | 15 | X(15) |  | N |
| Full\_height | Integer |  | (4)9 |  | Y |
| Full\_width | Integer |  | (4)9 |  | Y |
| Updated\_at | STRING | 36 | X(36) [UUID] |  | Y |
| Uuid | Numeric | 2 | -.999999 |  | Y |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Entity: **Menu** | | | | |  |
| Field Name | Type | Precision | Format | Key | Null |
| id | INTEGER |  | (10)9 | PK | N |
| name | STRING |  | (10)9 | FK | N |
| sponsor | STRING |  | ----.99 |  | Y |
| event | STRING | 15 | X(15) |  | Y |
| venue | STRING |  | (4)9 |  | Y |
| place | STRING |  | (4)9 |  | Y |
| physical\_description | STRING | 36 | X(36) [UUID] |  | Y |
| occasion | STRING | 2 | -.999999 |  | Y |
| notes | STRING |  |  |  | Y |
| call\_number | STRING |  |  |  | Y |
| keywords | STRING |  |  |  | Y |
| language | STRING |  |  |  | Y |
| date | DATE |  |  |  | Y |
| location | STRING |  |  |  | Y |
| location\_type | STRING |  |  |  | Y |
| currency | STRING |  |  |  | Y |
| currency\_symbol | STRING |  |  |  | Y |
| status | STRING |  |  |  | Y |
| page\_count | INTEGER |  |  |  | Y |
| dish\_count | INTEGER |  |  |  | Y |

## Data Quality Initial assessment

For an initial assessment, following data quality checks were performed and respective violations were listed.

**Dish.csv file:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Issue type** | **Description** |
| Name | Case Standardization | The name does not appear in the standard case. Some names are upper-case, some lower-case and other mixed-case.   * There are 426985 distinct dishes * There are 398443 distinct dishes case insensitive   e.g.  Dish.id = 1 , Dish.name = 'Consomme printaniere royal'  Dish.id = 397198, Dish.name = 'Consomme Printaniere Royal' |
| Name | Extra Spaces | There is extra space   * Leading dish names * Training dish names * In between dish names   e.g.  Dish.id = 131274, Dish.name = 'Consomme printaniere royal'  Dish.id = 397198, Dish.name = ' " " kidneys' |
| Description | Completeness | The entire field is empty and unusable |
| Name | Extra quotes | There are extra quotes in dish names  e.g.  Dish.id = 1788, Dish.name = 'Veuve Clicquot "Yellow Label"' |
| Name | Invalid characters | There are invalid characters like (!,@,#,{ etc.) in dish names  e.g.  Dish.id = 2839, Dish.name = 'E. & J. B. \*\*\*' |
| menus\_appeared | Plausibility | There are 2412 Dishes with menus\_appeared = 0  The field may be re-calculated from references |
| menus\_appeared | Correctness | There are differences in menus appeared and actual menu count for dish  e.g.  id =19, menu\_appeared = 16, actually appeared = 15 |
| times\_appeared | Plausibility | Several 0 or negative values  -- 1 Dishes appeared -10 times ??? [MIN]  -- 11900 dishes appeared 0 times !!!  -- 372 dishes appeared 19 Menus [MAX] |
| times\_appeared | Correctness | There are differences in times appeared and actual count for dish in menus  e.g.  id =17, times\_appeared = 535, actually appeared = 536 |
| First\_appreaed | Correctness | Many dishes have first appearance earlier than menu date |
| First\_appreaed | Plausibility | Several dishes have first\_appeared year later than last\_appreared year |
| Lowest\_price | Correctness | Dish lowest price should not be negative. Several dishes have 0 lowest price. It may not be an issue but worth analysing especially when nulls are allowed. |
| Highest\_price | Correctness | Dish lowest price should not be negative. Several dishes have 0 lowest price. It may not be an issue but worth analysing |
| name | Duplications | Same standardized dish name has multiple entries with different context |

**MenuItem.csv file:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Issue type** | **Description** |
| dish\_id | Null as FK,  lineage broken | 241 menu items have no value for dish id  e.g.  menu\_item.id = 19171 , Menu\_item.dish\_id = NULL |
| dish\_id | Referential integrity | 3 dish ids in menu item which do not exist in dish  e.g.  menu\_item.id = 619133 , Menu\_item.dish\_id = 220797 |
| Price | Completeness | More than 446K menus have null in price. It may be overwritten by average dish price from   * Highest price in menu item (58 cases) * Menu Items for same dish * Dish lowest and highest prices |
| high\_price | Completeness | More than 1.2M menu items have no high price. It may be overwritten from  Lowest price in menu item (~800k cases)  Corresponding dish highest prices |
| Price | Correctness | in 1278 cases, High\_price is strictly less than price  e.g.  MenuItem.id = 1455, price = 40, high\_price = 0.4 |
| Created\_at | plausibility | in 2874 cases updated\_at timestamp is earlier than created\_at timestamp |
| xpos | Correctness | xpos value is 0 in 9 cases. May be correct but worth a look. |
| ypos | Correctness | ypos value is 0 in 1 case. May be correct but worth a look. |

**MenuPage.csv file:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Issue type** | **Description** |
| id | Referential integrity | 40334 Menu pages ar not referred by any menu items. This may not be a problem (e.g. title page etc.) but worth analysing. |
| menu\_id | Referential integrity | 5803 Menu ids in menu page which do not exist in Menu file  e.g.  menu\_page.id = 119 , Menu\_Page.Menu\_id = 12460 |
| page\_number | Plausibility | Range of Page\_numberis (1,74)  Is 74 pages long menu plausible? |
| page\_number | Completeness | Missing 1202 values. Can be constructed from image\_id and menu\_id |
| Image\_Id | Correctness | 23 image\_ids are alphanumeric, rest are integers  e.g.  menu\_page.id = 24313 , Menu\_Page.image\_id = ‘ps\_rbk\_637’ |
| Image\_Id | Correctness | in 92 cases, menupage refers same menu and same panupage but different image\_id |
| full\_height | Completeness | in 329 cases, full\_height is null nut image\_id is known |
| full\_width | Completeness | in 329 cases, full\_width is null nut image\_id is known |
| uuid | Completeness | The entire field is empty and unusable |
| updated\_at | Format | the format of Updated\_at is that of a UUID, looks like fields UUID and Updated at are swapped |
| updated\_at | Uniqueness | Having UUID in it, this field has 2922 duplicated ids, several of these ids repeating as many as 10 times |

**Menu.csv file:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Issue type** | **Description** |
| name | Completeness | 14230/17423 rows are blank |
| name | Correctness | Among 3197 name values, only 793 are unique |
| name | Consistency | There are values such as *[not given]* or *[restaurant name and/or location not given]*. Do these values sound reasonable when nulls/blanks are allowed and vice versa? |
| sponsor | Format & Case | The name does not appear in the standard case. Some names are upper-case, some lower-case and other mixed-case.  Punctuation of sponsor names is not standardized. e.g  R.M.S. Parthia R.M.S. "Parthia""" |
| sponsor | Completeness | 1561/17545 Missing Values |
| sponsor | Consistency | 57 records have ‘?’ as value  30 records have ‘[restaurant and/or location not given.]’  Are these reasonable when nulls/blanks are allowed and vice versa? |
| event | Format & Case | The name does not appear in the standard case. Some names are upper-case, some lower-case and other mixed-case.  e.g  MITTAGESSEN [DINNER] MITTAGESSEN - DINNER |
| event | Completeness | 9391/17545 Missing Values |
| event | Consistency | Similar or same values e.g  dinner, dinner/dinner, [dinner], daliy dinner, (dinner)  ?, <Blank> |
| venue | Format & Case | Names need to be standardized along with punctuation & cases e.g  ‘COMMERCIAL’, ‘COMM;’ |
| venue | Completeness | 9414/17545 Missing Values |
| place | Format & Case | Names need to be standardized along with punctuation & cases e.g  ZURICH, ZURICH, AUSTRIA  9422/17545 Missing Values |
| physical\_description | Completeness | 2777/17545 Missing Values |
| occasion | Format & Case | Names need to be standardized along with punctuation & cases e.g  SECULAR HOLIDAY; , SECULAR HOLIDAY |
| call\_number | Correctness | Many values do not represent a call number  e.g.  wotm, \_wotm, 1910-737 ItemA-D, Baratta 16 oversized (2 copies) |
| keywords | Completeness | All empty values |
| language | Completeness | All empty values |
| date | Correctness | Few Invalid Dates e.g 2928-03-26  586/17545 Missing Values |
| location\_type | Completeness | All empty values |
| currency | Completeness | 11089/17545 Missing Values |
| currency\_symbol | Completeness | 11089/17545 Missing Values |
| status | Completeness | Same value = ‘Complete’ in entire field |
| dish\_count | Correctness | in 214 cases, the value of dhis count is different to distinct dishes the menu can be connected to |

## Use cases

The usefulness of data can be judged by the potential use case it may serve. The raw data has a number of issues as listed above but may still be valuable for several analytical scenarios and use cases. However once cleaned, further uses can be foreseen, some of them listed below:

* + 1. **Fitness for use as is**

Although the data is not clean enough for many useful use cases, still here are some.

* The data may be used by business owners to generate menus and see historical variations in the dish listings.
* The data may be used to popular dishes based on the number of times a dish is listed. It may also be used to find dishes previously listed but not offered anymore as a criterion of the unpopularity of the dish.
* The data may be used for building a New York Menus search engine where users can search for restaurants and the dishes associated with them.
* The data may be used to search businesses (with call number) which offer a specific dish
* One can analyze how menus have changed across time in terms of their page size and number of dishes.
  + 1. **Fitness for use after data wrangling**

Once the initial quality issues have been addressed the data will be fit for the following and other similar use cases:

* All use cases for which the data may be used without cleaning may be served with much higher quality after the data is cleansed.
* Once gaps in the data are filled and formats are standardized, Machine Learning models may be applied to uncover patterns in dish offerings, prices and locations.
* Once correct chronology of menu listings may be done, the dataset may be used to study how eating preferences evolved with time by looking at how the popularity of dishes changed across time.
* Once duplication in dishes are merged, upselling, cross-selling and competitive pricing analysis for dishes across locations may be performed.
* If a review dataset (such as yelp reviews) may be combined with this data (based on cleansed business names etc.), rating or sentiment analysis of dishes may be performed per restaurant.
* If the dataset may be mapped with geo-coordinates, location-based dish and menu analysis may be performed.
* A recommender system may be used to recommend dishes in a price range or location.