CS513 Final Project
Phase 1 Report

Contributors:

Roy - ralda2@illinois.edu
Atif Malik - atifqm2@illinois.edu
Dan Crosson - dfc3@illinois.edu
Jason Lundquist - jasondl3@illinois.edu

Introduction

This document is the Phase 1 project report for a data cleaning process to be performed on New York Public Library's historical menus dataset. The report outlines the proposed data cleaning, the details of the steps to perform and the methods for measuring the data cleaning outcome. The New York Public Library's restaurant menu collection holds data about menus and dishes from 1840 to present. This is a crowdsourced dataset collected through spreadsheets and APIs. Since the data is crowdsourced and collected via various means, the data quality is very poor.
The objective of this project is to apply various data cleaning techniques and analyze their effectiveness on the data set. After performing data cleaning, it is expected that the data will be "fit for use" in the data analytics use cases outlined in this report.

Data Source:

NYPL_DataSet (New York Public Library's crowd-sourced historical menus dataset)

Data Set Description:

A crowd-sourced dataset of restaurant menus over time. The dataset contains 4 csv files: Menu, Dish, MenuPage, and MenuItem. Dish, MenuPage, and MenuItem contain statistics and ancillary information about the menus in the Menu file. The Menu file contains information about all the menus in the dataset including the restaurant name, occasion, meal, location, length, and physical description.

Initial Quality and Problems:

The Menu file is riddled with missing information and misspellings. The other three files also have some missing data, but not as much as the Menu file.

Abstract:

Report containing proposed data cleaning to be conducted on New York Public Library's crowd-sourced historical menus dataset with a goal to curate the datasets for data analytics use cases.

Use Cases:

1. Dish price and other details.
2. Demand for dishes over time and contributing factors.
3. Ingredients of dishes and their corelation to prices
4. Correlation of demand with price, location and other factors.
5. Price and location changes over time.

Technology:

YesWorkflow
Python
OpenRefine
SQLite

Data

Dish.csv
This file contains the names of the dishes along with high and low prices, when the dish first appeared in a menu, when it last appeared and in how many menus it appeared in.

Menu.csv
This file contains details about the menus, including the name, place, occasion, event, location, venue, sponsor. It also contains supporting information, number of pages in the menu, number of dishes in the menu, language of the menu, currency of payment and other descriptive fields.

MenuPage.csv
This data is related to the menu file data and provides details of menu pages. It provides the image of the menu page, height and width of the page and page number, and the menu id from the menu file.

MenuItem.csv
This data is related to the menu, dish and menu page files and provides the information related to each menu item. For each menu item, it provides the dish id from dish.csv and menu page id from menupage.csv. It also provides other details of menu items such as price, creation type, and the location coordinate of the place.

Data Quality

We will focus on the following key areas for data quality assessment:

- Leading/trailing spaces
- Additional spaces
- Inconsistent case
- Inconsistent data - abbreviations and full text forms present in the same columns
- Inconsistent date format for date columns
- Blank values for columns
- Inconsistent representation or missing information
- Special Characters like #%?\()[], Carriage Return and Line Feed, and TAB characters
- Spelling
- Inconsistent value format
- Double quote usage
- Key values mismatch between parent/child entities
- Negative or incorrect values for numeric data such as price
- Data integrity issues
- Poor formatting
- Ambiguity