PROJECT : DATA PREPARATION FOR FOOD INSPECTION

SD6104 DATA PREPARATION

TOTAL MARKS: 100 (40% weightage)

Due Date of Report: April 21, 2025; 11:59 PM

Consider the food inspection dataset at https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/about_data. It contains 17 columns and 287K rows. The dataset is derived from inspections of restaurants and other food establishments in Chicago from January 1, 2010. Inspections are performed by staff from the Chicago Department of Public Health's Food Protection Program.

Your goal is to preprocess the data to understand their characteristics and then store it in a relational database for subsequent analytics. You can assume the specific analytics tasks you wish to perform. The details related to the attributes in the dataset are described in the website. A copy of the dataset is also uploaded in the project folder.

PROJECT TASKS

Specifically, you need to undertake the following tasks:

- Profile the data sets (single- and multi-column) to understand and visualize the characteristics of the datasets and relationships between the columns. Specifically, you should perform the following tasks:
    - Single-column profiling to gather various statistics and distributions of the columns.
    - Find correlation between columns using association rule mining technique.
    - Discovery functional dependencies from the columns.
    - Find inclusion dependencies between the columns.
    - Devise techniques to visualize the results of the aforementioned data profiling tasks. Explain what useful information you have garnered by performing the aforementioned tasks.
    - Find data quality problems in the data using techniques taught in your course. How can you address them?

SOFTWARE REQUIREMENTS

You should use Python to implement data preparation tasks. All program should run on an Windows machine. You are free to use any off-the-shelf packages.

SUBMISSION REQUIREMENTS

You submission should include the followings:

- **Software:** Please submit a single python file for ease of checking. You should also provide a read-me file to describe how your software needs to be run.
- **Presentation:** Each group shall present their solution on **Week 14/15**. Your presentation should include a demonstration to show how you have achieved the above tasks. Details related to presentation will be posted nearer to the date.
- All submissions will be through NTU Learn. The submission site will be opened closer to the deadline.

GRADING POLICY

The software carries **50%** of the marks. The remaining **50%** will be for the presentation. You will be graded on:

- Diversity of data preparation tasks implemented.
- Efficient design and implementation of various tasks.
- Visualization of results of data preparation.
- Successful and correct execution of the code.
- Quality of the presentation.
- Demonstration of understanding of data preparation concepts.
- ***Note: Late submission will be penalized. You will lose 10 marks/day.***