CS 513 Data Cleaning Group Project Phase-I

Team 65

Wei Dai <weidai6@illinois.edu>

Xiaoyu Wen <xwen20@illinois.edu>

Felicia Liu <liu318@illinois.edu>

1. Identify a dataset

The Chicago food inspection dataset for this project is retrieved from Kaggle: https://www.kaggle.com/datasets/chicago/chi-restaurant-inspections. This dataset was collected by the City of Chicago Department of Health. This dataset is about the inspection results, date, violations, and location information of various types of retail food establishments in Chicago including restaurants, grocery stores, bakeries, convenience stores, hospitals, nursing homes, day care facilities, shelters, schools, and other temporary food service events. This data covers inspections from 01/02/2013 to 08/28/2017. This dataset has 153,810 records, and each record has 18 attributes. We will perform data cleaning to this dataset to support our use cases in this project.

2. Develop use cases

2.1 Target Use Case U1

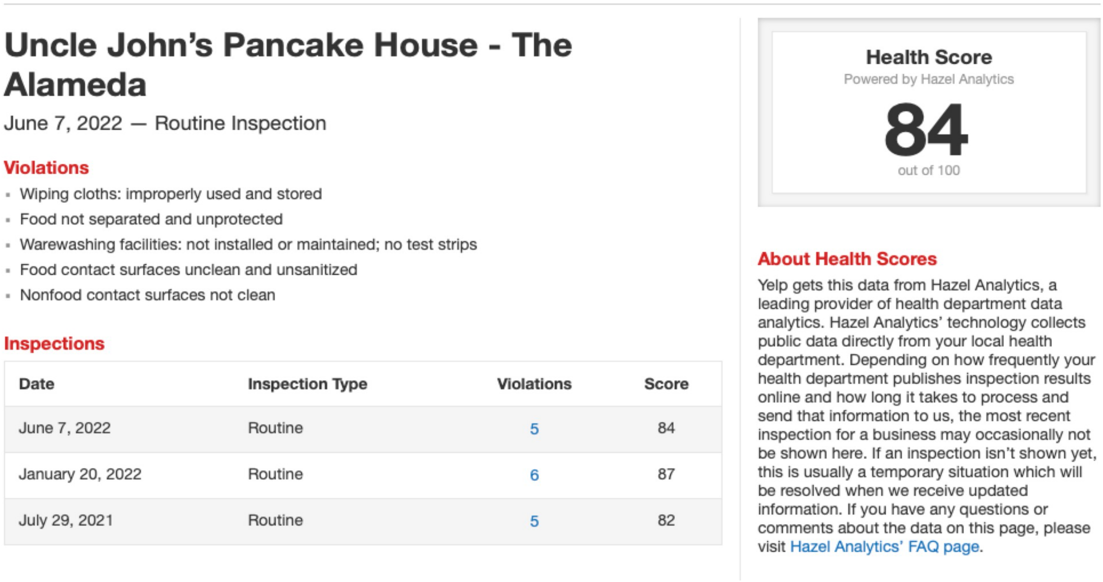
(1) Scenario 1: **Use cleaned data to analyze the relationship between facility's location and inspection results/business status.**

Inspection results: A family wants to move to Chicago. The quality of food is one of the key factors the family is concerned about. How could you help the family to know which community has better restaurants, grocery stores, and mobile food dispensers? Explanation: If the inspection pass rate in each ZIP code area is displayed as color over the Chicago map, it could provide users the food quality from different types of retail food establishments in each area.

Business status: An investor intends to enter the food market in Chicago, and is eager to understand which location he/she should select. Can you provide some insights to him/her leveraging the Chicago food inspection dataset? - Explanation: By glancing at the Chicago food inspection dataset, the facility's location seems to have a critical impact on the food business status (operating or out-of-business) in Chicago because the dataset contains columns “Address”, “City”, “State”, “ZIP”, “Latitude”, “Longitude”, “Location”, however, it is difficult to draw a conclusion before cleaning the dataset. For instance, if we combine the “location” information of each record with a map, it could provide us some insights on the relationship between the facility's location and business status in Chicago.

(2) Scenario 2: **Use cleaned data to develop a web application that provides health scores for customer usage.**

This web application is inspired by Yelp ‘’Health Score” feature. The main idea is that, when users search for facilities such as restaurants, daycares, schools, this application provides health information for users to consider. This information contains inspection date, inspection type and the violation numbers and details for each violation. Then based on the risk level (high, medium, low) of each violation, this application provides a final score for each facility. Also, users can use multiple ways to search the facility: by name, by address or by clicking on the map — since latitude and longitude data is also provided in this dataset. The reasons we need sufficient data cleaning for this usage are as follows: first, some columns are not needed, such as license number, and the inspection results; and the ‘location’ column is the combination of latitude and longitude, just one of them may be used. During the cleaning procedure, those data may be removed directly. Second, there are many duplicate addresses. The reason is that some facilities are already out of business, so the new facility may use the same address. Then the old facilities need to be removed from the dataset, since it does not exist any longer.



Uncle John's Pancake House - The Alameda

June 7, 2022 — Routine Inspection

Violations

- Wiping cloths: improperly used and stored
- Food not separated and unprotected
- Warewashing facilities: not installed or maintained; no test strips
- Food contact surfaces unclean and unsanitized
- Nonfood contact surfaces not clean

Inspections

Date	Inspection Type	Violations	Score
June 7, 2022	Routine	5	84
January 20, 2022	Routine	6	87
July 29, 2021	Routine	5	82

Health Score

Powered by Hazel Analytics

84

out of 100

About Health Scores

Yelp gets this data from Hazel Analytics, a leading provider of health department data analytics. Hazel Analytics' technology collects public data directly from your local health department. Depending on how frequently your health department publishes inspection results online and how long it takes to process and send that information to us, the most recent inspection for a business may occasionally not be shown here. If an inspection isn't shown yet, this is usually a temporary situation which will be resolved when we receive updated information. If you have any questions or comments about the data on this page, please visit Hazel Analytics' FAQ page.

Figure 1. An example of Health Score from Yelp

2.2 Minor Use Case U0

(1) Scenario 1: **Leverage the original dataset to understand the correlation between facility's type and inspection results.**

Inspection Use Case: The Chicago government hopes to find out whether the different facility types affect the food inspection results (pass or fail) in Chicago. Can you provide some insights to the government leveraging the Chicago food inspection dataset?

Explanation: The facility types in the dataset are divided as 399 different types, and for each type category, it is easy to identify the food inspection result (pass or fail). As a consequence, it is not necessary to perform data cleaning since the original dataset is capable of granting sufficient information for analysis.

(2) Scenario 2: **When users would like to check whether the facility is inspected, the original data is already very clean.**

In the web application usage scenario mentioned above, one corner case is that if the application user only needs to check whether the facility has been inspected or not, or the date this facility is inspected, then the data is ready to use. Because the 'Inspection Date' column is already very neat. Date starts from Jan 4 2010 to Aug 28 2017, not any missing data at all. All we need is to change the data type from ‘object’ to ‘datetime’ in the Pandas library in Python.

2.3 Minor Use Case U2

(1) Scenario 1: **Performing data cleaning is not sufficient to come to a conclusion that a certain type of violation will cause failure of food inspection.**

Inspection Use Case: The Chicago government wants to understand all kinds of violations in the food inspections and whether a certain type of violations will definitely lead to the failure of food inspection in Chicago. Can you provide some insights to the government leveraging the Chicago food inspection dataset?

Explanation: The Chicago food inspection dataset contains some useful information like “Violations”, “Results”, however, the violation types are too difficult to be categorized into measurable types. Hence, performing data cleaning is insufficient to identify the correlation between the violation types and failure of food inspection in Chicago.

(2) Scenario 2: **Performing data cleaning is never sufficient to provide a final choice for users.**

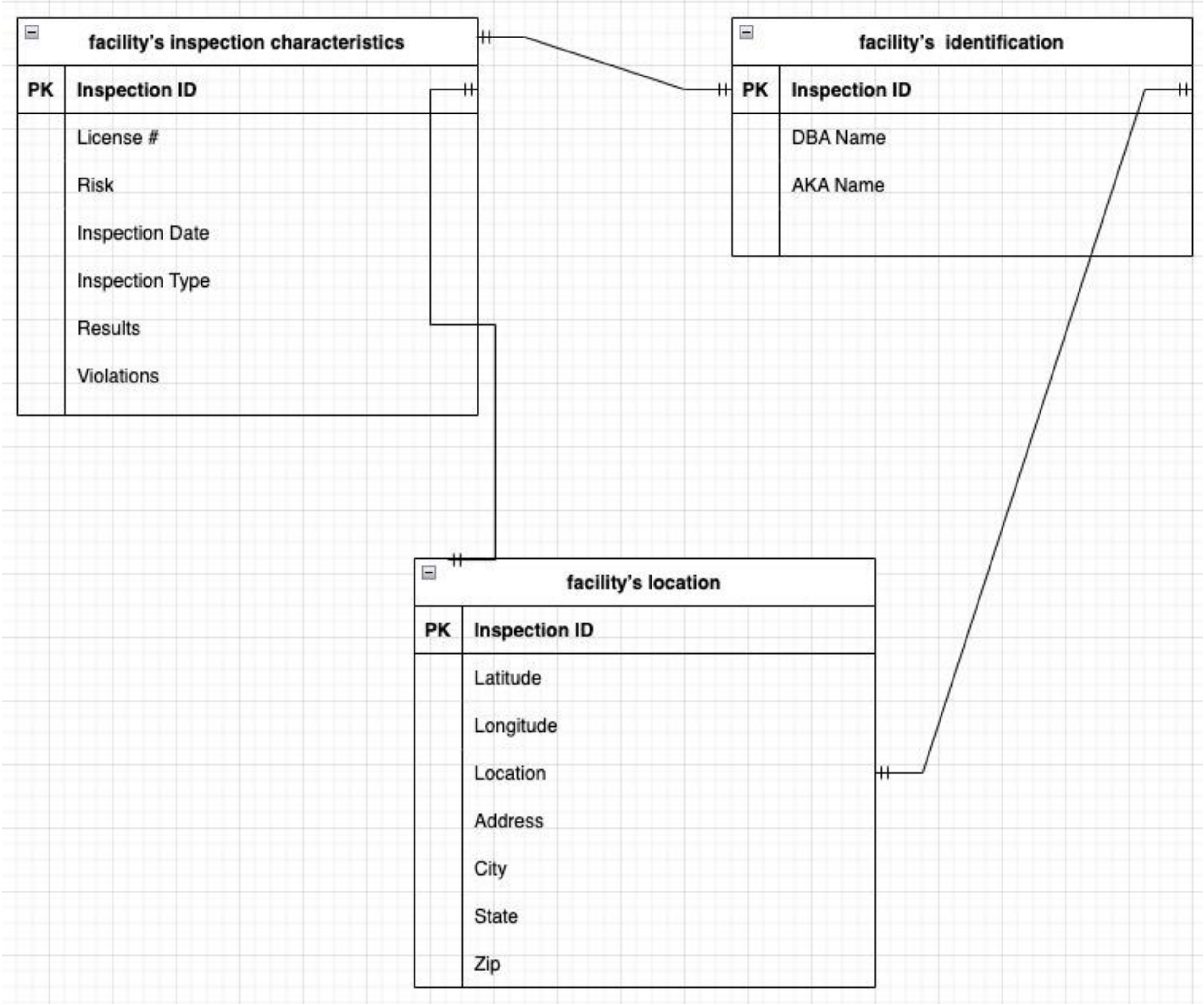
In the web application usage scenario mentioned in the U1 case. If the user would like to make the final decision of whether to choose this facility or not, then the dataset is not very helpful. This dataset can provide details of each violation and the risk level and give the final scores based on the weighted of each level. However it is the user’s choice to go to any of this facility.

3. Describe the dataset D.

The dataset’s size is 184.8M with 153,810 entries and 17 columns. Overall, the dataset provides information regarding the facility’s inspection characteristics, identification and location. For further detail, the below table and graph granted a brief description in each column of the dataset and ER Diagram of this dataset.

Name	Type	Description
Inspection ID	int	Unique inspection ID of the facility
DBA Name	String	Name of the facility
AKA Name	String	Name of the facility
License #	int	License Number of the facility according to Chicago food inspection dataset
Facility Type	String	Facility types of the facility
Risk	String	Risk rating of the facility
Address	String	Address of the facility according to Chicago food inspection dataset
City	String	City of the facility according to Chicago food inspection dataset
State	String	State of the facility according to Chicago food inspection dataset
Zip	int	Zip Code of the facility according to Chicago food inspection dataset
Inspection Date	date	Inspection date of the facility
Inspection Type	String	Inspection type of the facility
Results	String	Inspection result of the facility
Violations	String	Inspection violations of the facility
Latitude	float	Latitude of the facility
Longitude	float	Longitude of the facility
Location	float	Location including latitude and longitude information of the facility

Table 1 - Brief description of the dataset



Graph 1 - ER Diagram of the dataset

4. **List obvious data quality problems**

(1). There is a lot of missing data. For example, there are many Licenses #0, which is obviously incorrect; some facilities' risk levels are left blank. Some facilities' risk levels are high, whereas their violations are left blank.

(2). Many facilities are using the same license numbers. There may be some duplicates, or one facility provides different services. These problems are left to further exploration.

(3). Some column data types are incorrect. For example, column ‘zip’ and ‘License #’ should be converted to integer. ‘Latitude’ and ‘Longitude’ should be converted to decimal. This is a very minor problem that can be fixed very quickly by using tools mentioned in class.

(4). Many facilities are using the same address. There are many reasons, one main reason is that some facilities are out of business, and new facilities are using the same address. This is reasonable, and we are using this part of data for further analysis. However, some existing facilities are sharing the same address. This may be a crucial task for our further exploration.

(5). Many facilities are using the same name. Some are chain stores, the others may be duplicates or exactly different facilities using the same name.

(6). Column ‘Location’ is the combination of column ‘Latitude’ and ‘Longitude’, just one of them needs to be used in our analysis.

(7). Many facility types can be clustered. For example, ‘candy maker’ ‘candy shop’ and ‘candy store’.

(8). In the ‘facility type’ column, some facilities provide multiple services such as ‘bakery/restaurant’ and ‘bakery/deli’.

(9). ‘Inspection type’ column can be clustered. For example, ‘task force for liquor’ and ‘TASK FORCE LIQUOR’ can be combined.

5. Initial plan

S1: List all the columns of the dataset in Python pandas dataframe and understand the meaning of each column.

S2: Determine which columns are most useful to the target use case U1 and minor use cases U0, U2. Determine what is the best data type of these useful columns to support our use cases. Check which columns have the improper data type. Check which entries have missing value, typo, and other anomalies.

S3: Use python to remove the irrelevant columns, delete duplicate entries, drop entries with missing or abnormal values, correct typos, and convert data type according to what use cases needs.

S4: Check if any duplicate entries, missing values, anomaly data, or improper data types still exist. If any of these problems stil exist, improve our date cleaning method and repeat the S3 and S4 until the dataset is well cleaned. Perform data visualization (Choropleth map with Tableau) to support use cases.
S5: Documenting the changes made to the dataset. Documenting our data cleaning procedure.

Assignment of tasks:

Week 1. Python: Xiaoyu, Wei, Felicia
Week 2. Tableau: Wei, Xiaoyu
Week 3. Workflow: Felicia
Week 4. Final report: Xiaoyu, Wei, Felicia