CS 513: Theory & Practice of Data Cleaning Final Project
University of Illinois at Urbana-Champaign - Summer 2023
Phase 1 Report

Abstract
This report covers our initial understanding and plan on the dataset we chose for our final project. In this project, we will perform data cleaning through various tools and techniques that we learned in the course.

Team Details
Team-ID: 120
Members:
- Jeremy Ahn (jcahn2@illinois.edu)
- Nithin Nathan (nnatha3@illinois.edu)
- Ratul Saha (ratuls2@illinois.edu)

1. Dataset Chosen

We chose the Chicago Food Inspection dataset for our project.

This dataset can be found at the following link:
https://uofi.app.box.com/s/whvfh9jio38ck0m9gz58s31srx8iwg4i/folder/159094327937

The original source for this dataset can be found at the following link:
https://www.kaggle.com/datasets/chicago/chi-restaurant-inspections

2. Description of Dataset

The provided dataset contains inspection details of around 24,000 facilities located in Chicago over the course of 8 years (2010-2017). The business details contain the business name, legal name, address details, and facility type. Location is identified by address, city, state, zip code, latitude and longitude. Inspection details capture inspection date, observation, risk type and result.

We are planning to arrange the data of inspection details into 3 files, as described below:

1. Business: This file contains business identifiers, business names and facility type for every business listed. We will consider this as a master dataset.
2. Location: This file contains all the locations and their associated details, irrespective of which facility or what business is operating from there. This file contains geographical location and address details. We will consider this as another master dataset.
3. Inspection: This file contains all the inspection details performed on the businesses. We will consider this file as a relationship/fact dataset.

Our objective is to clean and improve the data quality of the dataset to establish the following constraints:

- Every License # in the Business file defines a unique business.

- Every Location ID in the Location file defines a unique geographic location.
- Every Inspection ID in the Inspection file defines a unique inspection case.

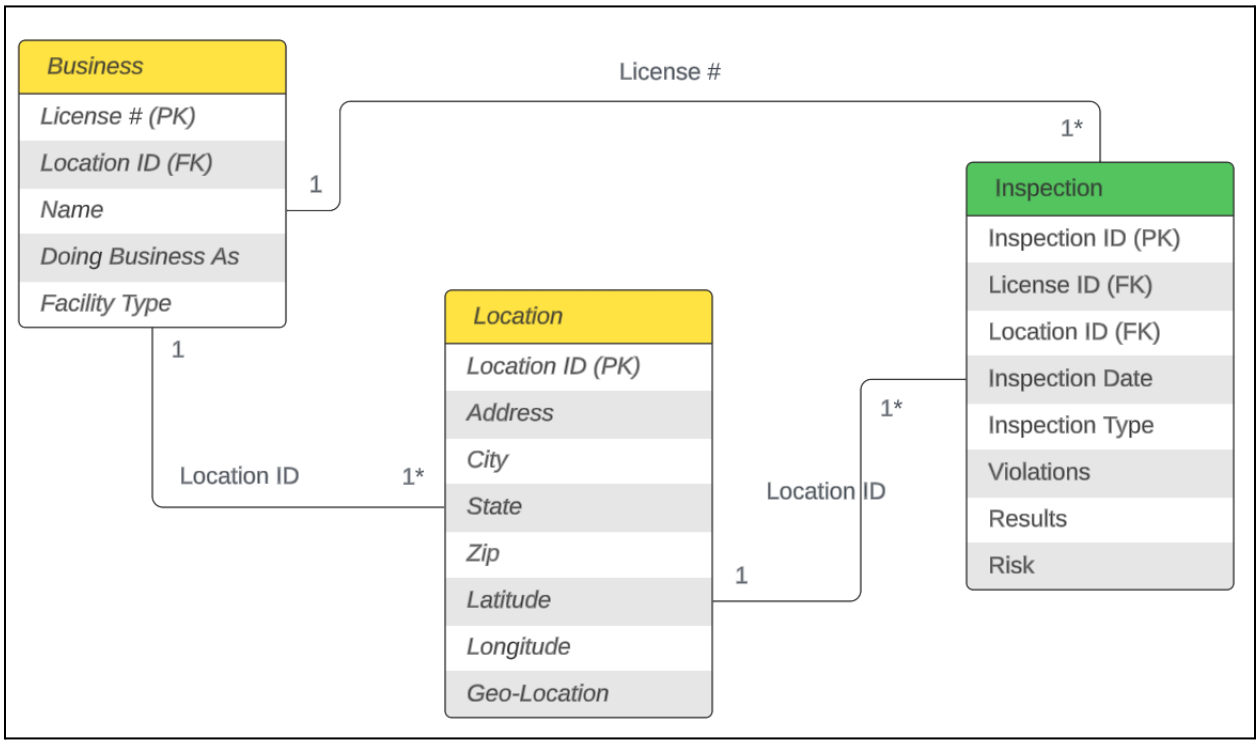
By establishing primary and foreign key constraints, we can also ensure the following data quality parameters:

- Every Location ID present in the Location file is also present in the Business file.
- Every License # present in the Inspection file, is also present in the Business file.

Upon further normalization, we can identify the following real-life situations, integrity constraints, functional dependencies:

- One business can have more than one inspection case, meaning that it could have been inspected multiple times.
- One business can have more than one location, meaning that the business could have moved to different locations.

The Entity Relationship Diagram is explained below.



Each of the data fields are described below:

- Inspection ID: Unique ID for each inspection case performed on a business at a location.
- Doing Business As (DBA) Name: The name of a business doing business as or also known as legal/registered name.
- AKA (Also known as) Name: Common name of the business, can be the same as DBA or blank.

- License #: Unique licensing number assigned by the Department of Business Affairs and Consumer Protection
- Facility Type: Type of the business facility.
- Risk: Each establishment is categorized as to its risk of adversely affecting the public's health, with 1 being the highest and 3 the lowest. The frequency of inspection is tied to this risk, with risk 1 establishments inspected most frequently and risk 3 least frequently.
- Address: The building name and street name
- City: City of the location. This field is blank in % of records. Also some data quality is noticed in this field.
- State: "IL" (2 digit state code of Illinois). This field is blank in % of records.
- Zip: 5 digit zip code of the facility. This field is blank in % of records.
- Inspection Date: Date when an inspection was performed.
- Inspection Type: Various purposes of inspection like license, complaint, canvass, expansion etc.
- Results: Final inspection outcome like pass, fail, not ready etc.
- Violations: The textual description of the inspection results. An establishment can receive one or more of 45 distinct violations (violation numbers 1-44 and 70).
- Latitude: The geographical position latitude of the facility.
- Longitude: The geographical position longitude of the facility.
- Location: The geographical longitude and latitude of the facility for locating the exact place.

More detailed description of the dataset is provided here
https://data.cityofchicago.org/api/assets/BAD5301B-681A-4202-9D25-51B2CAE672FF

3. Use Cases

Name	Zero Cleaning Use Case
Label	U0
System	Chicago restaurant and food establishment inspections (Jan 1, 2010 - present)
Actors	• Data analysts/Food inspection analyst • Chicago Food Inspection dataset (D) • Food Protection Division of the Chicago Department of Public Health (CDPH)
Goal	To study trends in the number of inspections made each day
Primary Actor	Data analyst/food inspection analyst
Preconditions	Data cleaning is *not* necessary
Basic Flow	Analyst performs a query on dataset D to count the number of entries

	associated with each unique date entry. Using the unique inspection IDs and the already consistently formatted date values, the analyst should be able to get at least a general idea of when and how many inspections were made in the Chicago area.
Alternate Flows	Alternatively, if the analyst needed to know the total number of Chicago food inspections made since 2010, a simple query count of all unique Inspection IDs should suffice to return the information without needing details of each inspection cleaned.

Name	Main Use Case
Label	U1
System	Chicago restaurant and food establishment inspections (Jan 1, 2010 - present)
Actors	● Data analysts/Food inspection analyst ● Chicago Food Inspection dataset (D) ● Food Protection Division of the Chicago Department of Public Health (CDPH)
Goal	To explore the effects of location on the risk evaluation of Chicago food inspections.
Primary Actor	Data analyst/food inspection analyst
Preconditions	Data cleaning is *necessary* and *sufficient*
Basic Flow	To explore the effects of location on risk evaluation, an analyst might seek to answer a question such as, “How does location affect the level of risk for a Chicago food establishment reported by an inspection?”. The analyst cleans data columns like ‘Risk’, latitude/longitude, location, and results by removing null entries, making the result values consistent, and quantifying the risk values. Once the data is sufficiently cleaned, the analyst performs queries to visualize and reveal patterns in the location and food inspection results.
Alternate Flows	Alternatively, instead of the latitude and longitude location, the analyst cleans the restaurant address and zip code columns, which could offer even more flexibility in terms of location specificity. By using the address or the zip code, the analyst is able to visualize both a specific and more regional grouping when used once again in combination with the risk results. We could think of employing an address data verification process e.g. USPS, to enrich and improve quality of address data.

Name	Never Enough Use Case
Label	U2
System	Chicago restaurant and food establishment inspections (Jan 1, 2010 - present)
Actors	● Data analysts/Food inspection analyst ● Chicago Food Inspection dataset (D) ● Food Protection Division of the Chicago Department of Public Health (CDPH)
Goal	To explore the prediction of inspection results based on the tone of the violation description's wording.
Primary Actor	Data analyst/food inspection analyst
Preconditions	Data cleaning is *not sufficient*
Basic Flow	The analyst spends hours combing each violation description and making sure each entry is consistent with spacing, formatting, etc. Each entry is stripped of filler words and categorized based on the positivity/negativity of each word. However, even with the data cleaned, the number of null entries and the conditionally passing category can make it difficult to make a comprehensive correlation. Given the size of the data and the limitations of data cleaning, this is quite impractical for an analyst to achieve.
Alternate Flows	The analyst uses neural networks and Natural Language Processing techniques to create a model that can predict the pass/fail results. Even with perfectly clean data, the model is not guaranteed to have 100% accuracy, and is simply not a task that data cleaning can achieve on its own in the first place. Either you need an alternate source to bring a good enough population to compensate for missing data through enrichment, or you can simply ignore the records with no violation description. In the later case, the analysis might not bring enough confidence to the outcome.

4. Data Quality Problems

Below are some data quality problems we identified. They are categorized below.

Null of blank values:

- License # is blank or zero, which could have been used to uniquely identify the business. However, blank value rows are unreliable for any analysis.

	Inspection ID	DBA Name	AKA Name	License #	Facility Type
44422.	1561809	ST. DEMETRIOS GREEK ORTHODOX CHURCH	ST. DEMETRIOS CHURCH		Special Event
72406.	1152076	ARGENTINA FOODS	ARGENTINA FOODS		Grocery Store
113447.	1214242	GOD'S BATTLE AXE PRAYER ACADEMY	GOD'S BATTLE AXE PRAYER ACADEMY		CHURCH/DAY CARE
124354.	521659	ST. DEMETRIOS GREEK ORTHODOX CHURCH	ST. DEMETRIOS CHURCH		Special Event
135568.	417318	AVALON COMMUNITY CHURCH/FREEDOM HOME ACADEMY			CHURCH/AFTER SCHOOL PROGRAM
2662.	2060022	CHICAGO BEST NAAN	CHICAGO BEST NAAN	0	Bakery
3210.	2059363	BBQ SUPPLY	BBQ SUPPLY	0	Restaurant
3719.	2050791	ST. EUGENE PARISH	ST. EUGENE PARISH Shaunnessy Center	0	Special Event
11807.	1979104	FIVE STARZ FOODS	FIVE STARZ FOODS	0	Grocery Store
15060.	1971047	PICNIC		0	RESTAURANT AND LIQUOR

- State is blank in a few cases.

isk	Address	City	State	Zip	Inspection Date
3)	1366 N MILWAUKEE AVE	CHICAGO		60642	07/02/2013
3)	1366 N MILWAUKEE AVE	CHICAGO		60642	06/11/2012
3)	1366 N MILWAUKEE AVE	CHICAGO		60642	06/05/2012

- Zip code is blank in a few cases.

Address	City	State	Zip
4 W JACKSON BLVD		IL	
4 W JACKSON BLVD		IL	
133 N WELLS ST		IL	
133 N WELLS ST		IL	

- Location (latitude/longitude) is blank in a few cases.

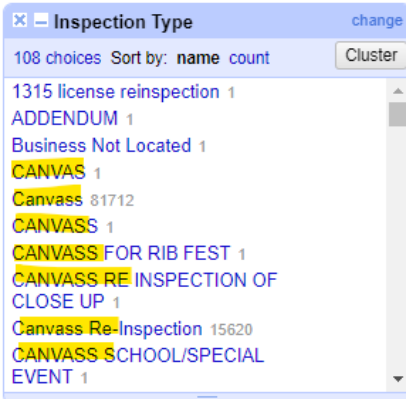
	violation code	Latitude	Longitude	Location
edit				
INAL PERLY MER ED AS nts: NSE THAT HES DES	30			
IG SOAP AND	12			

- The Address field has some values where the building number is a range. This might pose a problem as it will be difficult to locate the facility.

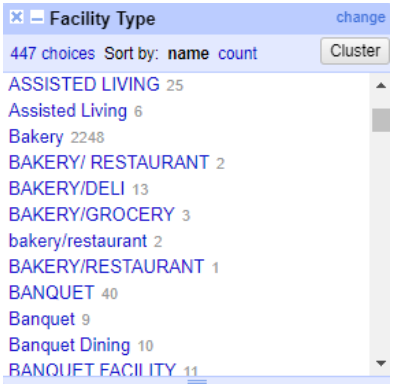
License #	Facility Type	Risk	Address	City	State	Zip
0	Long Term Care	Risk 1 (High)	1440-1448 E 75TH ST	CHICAGO	IL	60619

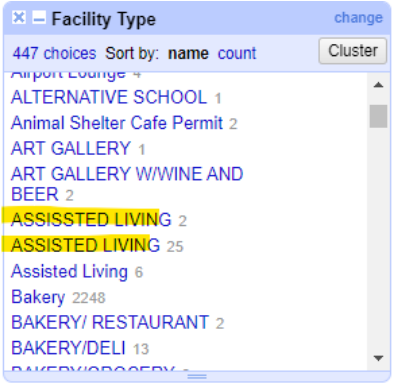
Non-Standard Data:

- Businesses with the same License # are named differently. This may present an issue when trying to consolidate businesses due to DBA Names being different for same License #'s.
- Inspection Type values are in different capitalization, sometimes with additional details. These values need standardization for better analysis and reporting.



- The Facility Type field has typos and non-standard values. The typos will need to be cleaned and the non-standard values will need to be addressed.

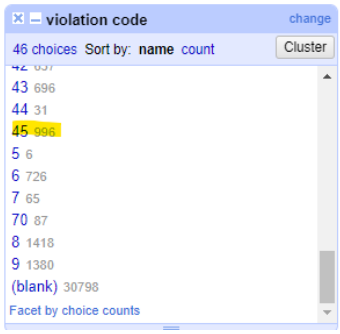




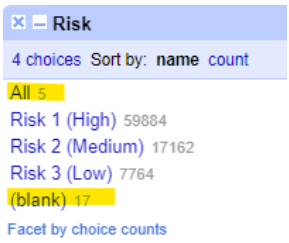
- City names are not in standard capitalization format. Some are concatenated, others contain typos, and a few are blank as well.

Invalid Data:

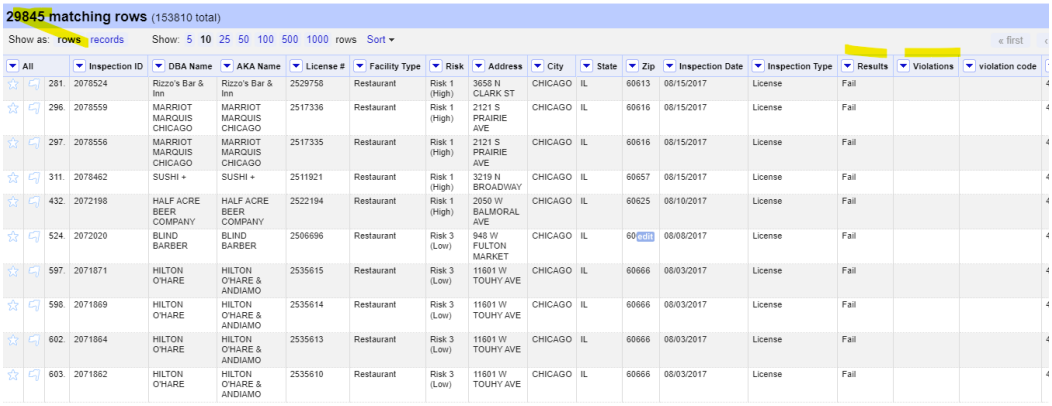
- Violation Codes can only be 1-44 and 70 (total 45 valid codes possible). There are 996 records with 45 as violation code.



- The Risk field has 2 values (blank and “all”).



- Violation text is not populated in Fail inspections cases.



29845 matching rows (153810 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 rows Sort « first

	All	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip	Inspection Date	Inspection Type	Results	Violations	violation code
	281.	2078524	Rizzo's Bar & Inn	Rizzo's Bar & Inn	2529758	Restaurant	Risk 1 (High)	3658 N CLARK ST	CHICAGO	IL	60613	08/15/2017	License	Fail		
	296.	2078559	MARRIOT MARQUIS CHICAGO	MARRIOT MARQUIS CHICAGO	2517336	Restaurant	Risk 1 (High)	2121 S PRAIRIE AVE	CHICAGO	IL	60616	08/15/2017	License	Fail		
	297.	2078556	MARRIOT MARQUIS CHICAGO	MARRIOT MARQUIS CHICAGO	2517335	Restaurant	Risk 1 (High)	2121 S PRAIRIE AVE	CHICAGO	IL	60616	08/15/2017	License	Fail		
	311.	2078462	SUSHI +	SUSHI +	2511921	Restaurant	Risk 1 (High)	3219 N BROADWAY	CHICAGO	IL	60657	08/15/2017	License	Fail		
	432.	2072198	HALF ACRE BEER COMPANY	HALF ACRE BEER COMPANY	2522194	Restaurant	Risk 1 (High)	2050 W BALMORAL AVE	CHICAGO	IL	60625	08/10/2017	License	Fail		
	524.	2072020	BLIND BARBER	BLIND BARBER	2506696	Restaurant	Risk 3 (Low)	948 W FULTON MARKET	CHICAGO	IL	60[edit]	08/08/2017	License	Fail		
	597.	2071871	HILTON O'HARE	HILTON O'HARE & ANDIAMO	2535615	Restaurant	Risk 3 (Low)	11601 W TOUHY AVE	CHICAGO	IL	60666	08/03/2017	License	Fail		
	598.	2071869	HILTON O'HARE	HILTON O'HARE & ANDIAMO	2535614	Restaurant	Risk 3 (Low)	11601 W TOUHY AVE	CHICAGO	IL	60666	08/03/2017	License	Fail		
	602.	2071864	HILTON O'HARE	HILTON O'HARE & ANDIAMO	2535613	Restaurant	Risk 3 (Low)	11601 W TOUHY AVE	CHICAGO	IL	60666	08/03/2017	License	Fail		
	603.	2071862	HILTON O'HARE	HILTON O'HARE & ANDIAMO	2535610	Restaurant	Risk 3 (Low)	11601 W TOUHY AVE	CHICAGO	IL	60666	08/03/2017	License	Fail		

- Functional Dependency between License # and DBA Name: License # does not derive DBA Name, in other words

```
(0.0, 'WHITE HEN PANTRY'): 1,
(0.0, 'WILLFEED PEACE KIDS CAFE'): 1,
(0.0, 'WONUT'): 2,
(0.0, 'unknown'): 1,
(241.0, 'HARD WATER BAR & GRILL'): 5,
(241.0, 'HOP HAUS ROGERS PARK'): 9,
(349.0, 'CLOVER CORP'): 3,
(349.0, "KELLY'S PUB"): 7,
(968.0, 'CAFE BERNARD'): 2,
(968.0, 'CHEZ MOI'): 9,
(1196.0, "FINN MC COOL'S / ALUMNI CLUB"): 5,
(1196.0, 'HOPSMITH TAVERN'): 8,
(1196.0, 'SNUGGERY / APARTMENT'): 2,
(1254.0, 'B AND J RESTAURANT INC'): 6,
(1254.0, "DINO'S PIZZA"): 9,
(1362.0, 'JEWEL FOOD STORE #3232'): 1,
(1362.0, 'MONTESSORI OF ENGLEWOOD'): 13,
(1578.0, '1000 LIQUORS / BIG CITY TAP'): 8,
```

```
In [7]:
1 print('<License # -> DBA Name> Functional Dependency Violations: '+str(len(chk5)))
2 print('Total Number of Records with <License # -> DBA Name> Functional Dependency: '+str(sum(chk5.values())))
```

```
<License # -> DBA Name> Functional Dependency Violations: 1339
Total Number of Records with <License # -> DBA Name> Functional Dependency: 6245
```

- An inspection can pass, pass with conditions or fail. Establishments receiving a ‘pass’ were found to have no critical or serious violations (violation number 1-14 and 15- 29, respectively). We found 17 (screenshot below) invalid violation codes (30-45, 70) which are present when Results were ‘Pass’.

	Results	Violation Code1
3571	Pass	70
13	Pass	45
107	Pass	44
105	Pass	43
83	Pass	42
239	Pass	41
32	Pass	40
320	Pass	39
24	Pass	38
18	Pass	37
27	Pass	36
3	Pass	35
68	Pass	34
69	Pass	33
22	Pass	32
219	Pass	31
14	Pass	30

Why and How Data Cleaning is Necessary to Support the Main Use Case (U1):
As per our use case, location is a directional factor to derive the regional clusters. Hence quality around longitude-latitude or address data is important. Incorrect location data might skew the regional clustering of the facilities. Hence if we try to find patterns of failures during inspections within a region, it might find some outliers that might not belong to that cluster. This in turn will give us wrong insight about failure reasons.

5. Phase 2 Initial Plan

Below is the high-level plan for Phase 2 of our project.

SL#	Activity	Responsibility	Timeline
S1	Review/update use case description and dataset description. ● Structure and content of the dataset. ● Derivation algorithms of use cases.	All	7/16
S2	Profile dataset to identify data quality problems ● Basic Data Profiling ● Detect Outliers ● Detect Errors ● Missing Value Analysis ● Discovery of Integrity Constraint Violations Data Quality Tools: ● OpenRefine ● Python ● YesWorkflow	Ratul	7/16
S3	Perform DC “proper” ● Perform SQL queries to explore dataset ● Use OpenRefine to perform data cleaning ● Use Python for further data cleaning as necessary ● Use YesWorkflow in conjunction with OpenRefine to create inner and outer workflow models	Jeremy	7/23
S4	Identify and perform quality improvements ● Remove Errors (Data Type, Data Values) ● Remove Special Characters and Spaces ● Clustering and Facet Operations ● Missing Value Operations ● Exclude Outliers	Nithin	7/27

	• Mark Integrity Constraint Violations		
S5	Document and quantify change • OpenRefine Recipe • SQL Queries • Python Script • YesWorkFlow Model • Comparison of Before and After Status of the Cleaning Operation • Establish the fact that the Main Use Case U1 can be performed optimally	All	7/29