CS 513 - Theory and Practice of Data Cleaning

University of Illinois at Urbana-Champaign

Prof. Bertram Ludaescher

Final Project - Phase 1

Team 194

Roberto Godoy <ard8@illinois.edu>

Steve McHenry <mchenry7@illinois.edu>

Fabricio Brigagao <fb8@illinois.edu>

July 1st, 2022.

1. Dataset Selection

For the proposed project, the **Chicago-Food-Inspection** dataset[1], which we refer to as *D*, was selected. Of the datasets that we reviewed, it appears to have several possible interesting use cases in addition to ample opportunity for data cleaning.

2. Dataset Use Cases

2.1 Use Case U_1 – Primary Use Case

This dataset has great potential to provide deep insight into health department code enforcement history for a given food business, city-wide trends, and recent incidents of violations across the city of Chicago with a web application.

Hence, we define such an application's use case U_1 as the following:

> ***Provide the ability to query inspections based upon a logical disjunction of the following fields: Inspection date; Business name (both the business' legal name as well as its "also-known-as" name, if any); Business license number; Inspection result; Specific codes of violations.***

First, however, data cleaning is necessary to move the *D* into an improved state, *D'*, sufficiently fit for purpose in responding to such requests.

After the necessary data cleaning, sufficient to support our established use case U_1, the resulting data set *D′* could be used to provide aggregations such as:

- Select on a specific business license and observe its violations and corrections at the individual code level over a specified time interval.
- Search for all violations of a specific code (or codes) across the city within a given time interval, perhaps to discover and proactively educate operators of seasonal violation trends.
- Use the geographic coordinate data of each inspection to render failed inspections or code violations on a map of the city of Chicago, to identify regions of interest for potential risk and focused code enforcement activity.

[1] https://www.kaggle.com/datasets/chicago/chi-restaurant-inspections

2.2 Use case U_0 – No Data Cleaning Required

We define use case U_0, which can make immediate use of the dataset in its original state, without requiring data cleaning, as follows:

> ***Calculate the number of inspections which were fully performed vs. the number of inspections which were not fully performed for each month.***

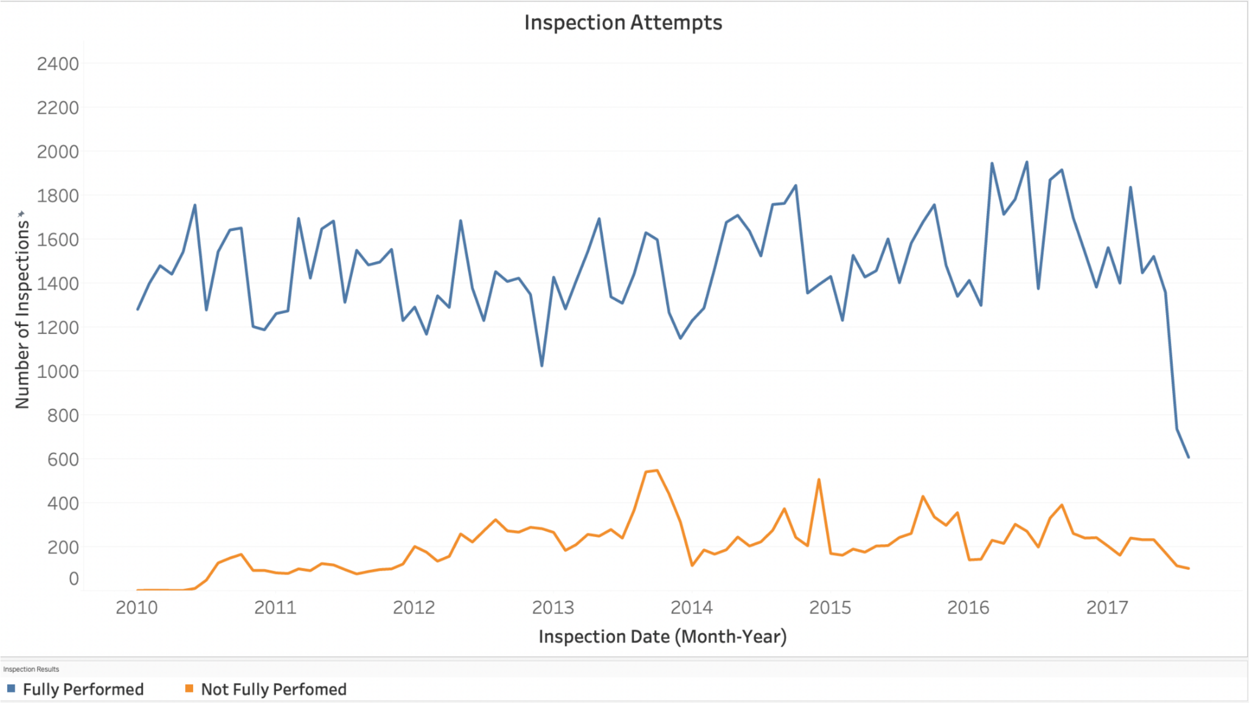
Where we define a fully performed inspection as an inspection during which the inspector was able to sufficiently inspect the business to determine the inspection result to be one of the following: “Pass”, “Pass w/ Conditions”, or “Fail”. In contrast, an inspection is considered not fully performed if the result is listed as one of the remaining possible options: “Business Not Located”, “Out of Business”, “Not Ready”, or “No Entry”.

U_0 can serve as a starting point to quantify the efficiency of the agency’s inspection attempt strategy during a given month. Ideally, the agency desires that all attempted inspections are fully performed to minimize inspectors’ unproductive time. However, issues and interruptions inevitably arise, such as being unable to locate the specified business, or the business being closed on that day.

The query to produce the result for use case U_0, provided below, was executed directly on the original dataset *D*. The dataset provided the “Inspection Date” field in “MM/DD/YYYY” format, and the “Results” field as a text field containing one of the 7 previously listed result values with consistent formatting (e.g., spelling, capitalization).

```
SELECT CONCAT(CAST(DATEPART(YEAR, FoodInspections.InspectionDate) AS
VARCHAR(4))
       ,'/'
       ,RIGHT('00' + CAST(DATEPART(MONTH, FoodInspections.InspectionDate) AS
VARCHAR(2)), 2)) AS YearMonth
      ,COUNT(CASE WHEN FoodInspections.Results IN ('Pass', 'Pass w/
Conditions', 'Fail') THEN 1 ELSE NULL END) AS FullyPerformedCount
      ,COUNT(CASE WHEN FoodInspections.Results NOT IN ('Pass', 'Pass w/
Conditions', 'Fail') THEN 1 ELSE NULL END) AS NotFullyPerformedCount
FROM FoodInspections
GROUP BY DATEPART(YEAR, FoodInspections.InspectionDate)
      ,DATEPART(MONTH, FoodInspections.InspectionDate)
ORDER BY YearMonth;
```

Below, we provide a graph, generated with Tableau directly from dataset *D*, showing the number of fully performed inspections in blue compared to the number of not fully performed inspections in orange from January 2010 through August 2017. Interestingly, it can be observed that, in general, especially near the beginning of the timeframe, the number of not fully performed inspections slowly yet consistently increased over time. In contrast, the number of fully performed inspections stays (very roughly) within the same range of approximately 1,250 to 1,750. Hence, it appears that over time the ratio of fully performed to not fully performed inspections has decreased. Perhaps the agency might consider adjusting their inspection attempt strategy to reduce unproductive time.

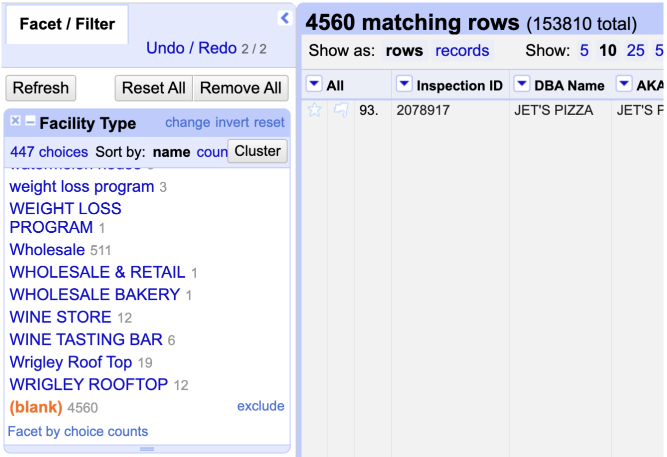


2.3 Use case U_2 – Data Cleaning is Not Sufficient

We define a use case U_2 which **cannot** be satisfied by neither *D* in its native state nor any cleaned form of *D*. U_2 is unsatisfiable because the data required is neither explicitly provided by *D* nor can it be derived from *D* (as *D'*). We define U_2 as:

> ***Calculate the number of failed inspections per facility type to predict which businesses may have the greatest health risk potential.***

This is not possible. First, 4,560 inspections (approximately 3% of the dataset) **do not specify the facility type** for the inspection, as the image below demonstrates.



Facet / Filter
Undo / Redo 2 / 2
Refresh
Reset All
Remove All

Facility Type
change invert reset
447 choices Sort by: name coun Cluster

weight loss program	3
WEIGHT LOSS PROGRAM	1
Wholesale	511
WHOLESALE & RETAIL	1
WHOLESALE BAKERY	1
WINE STORE	12
WINE TASTING BAR	6
Wrigley Roof Top	19
WRIGLEY ROOFTOP	12
(blank)	4560

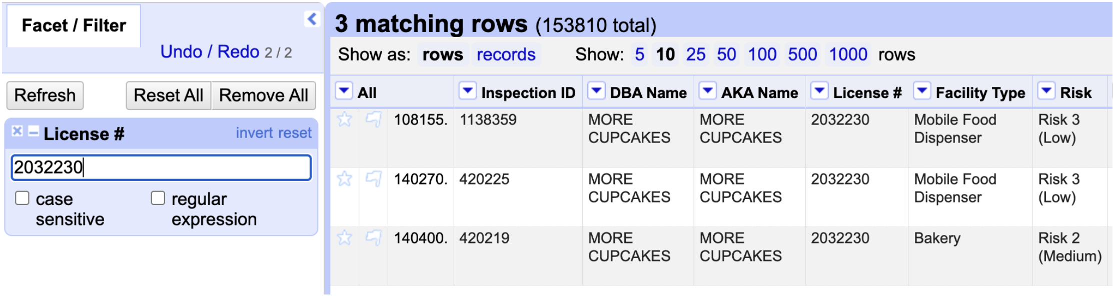
exclude
Facet by choice counts

4560 matching rows (153810 total)
Show as: **rows** records Show: 5 **10** 25 5

All	Inspection ID	DBA Name	AKA
93.	2078917	JET'S PIZZA	JET'S F

Furthermore, although one might attempt to infer the corresponding facility type based upon other inspections for the same business, many establishments which lack a facility type for one inspection also lack a facility type for all other inspections.

Secondly, some businesses are listed in the dataset with **conflicting facility types** at different inspections. For example, the business with **DBA Name** “*MORE CUPCAKES*” and **License #** “2032230”, had three inspections. The first inspection listed its facility type as “Bakery”, but the other two listed its facility type as “Mobile Food Dispenser”.



Facet / Filter
Undo / Redo 2 / 2
Refresh
Reset All
Remove All

License #
invert reset
2032230
☐ case sensitive
☐ regular expression

3 matching rows (153810 total)
Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows

All	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk
108155.	1138359	MORE CUPCAKES	MORE CUPCAKES	2032230	Mobile Food Dispenser	Risk 3 (Low)
140270.	420225	MORE CUPCAKES	MORE CUPCAKES	2032230	Mobile Food Dispenser	Risk 3 (Low)
140400.	420219	MORE CUPCAKES	MORE CUPCAKES	2032230	Bakery	Risk 2 (Medium)

Hence, given *D* or any improved dataset *D′*, we are unable to determine the correct interpretation and field values without seeking external guidance from the agency which produced the original data. In other words, since the facility type of a business can't be conclusively determined, we can neither use its past violations as criteria for the facility type risk calculation nor assign a predicted risk potential based upon the facility type.

3. Dataset Description

The Chicago Department of Public Health's Food Inspections dataset is a record set of the department's inspections performed between January 3, 2010 and August 27, 2017, totaling 153,810 inspection records.

Of these inspections, 153,494 (99.8%) were performed within the city of Chicago. However, a small number of surrounding cities also appear in the dataset, such as Naperville and Schaumburg.

Each record contains data about the corresponding inspection, such as the name and operating license of the business which underwent inspection, the address, the coded outcome of the inspection (e.g., "Pass", "Fail", etc.), and a concatenated list of coded violations that were encountered during the inspection.

Because several fields of the dataset are coded values – both singular and multi-valued – this dataset is a prime candidate for relational normalization into a database to both reduce redundant data and to enforce internal integrity constraints.

In the following table, we briefly enumerate and describe the fields from the original flat file representation, according to the official specification[2] provided at the Chicago Data Portal, to show how the source dataset maps onto our proposed database schema.

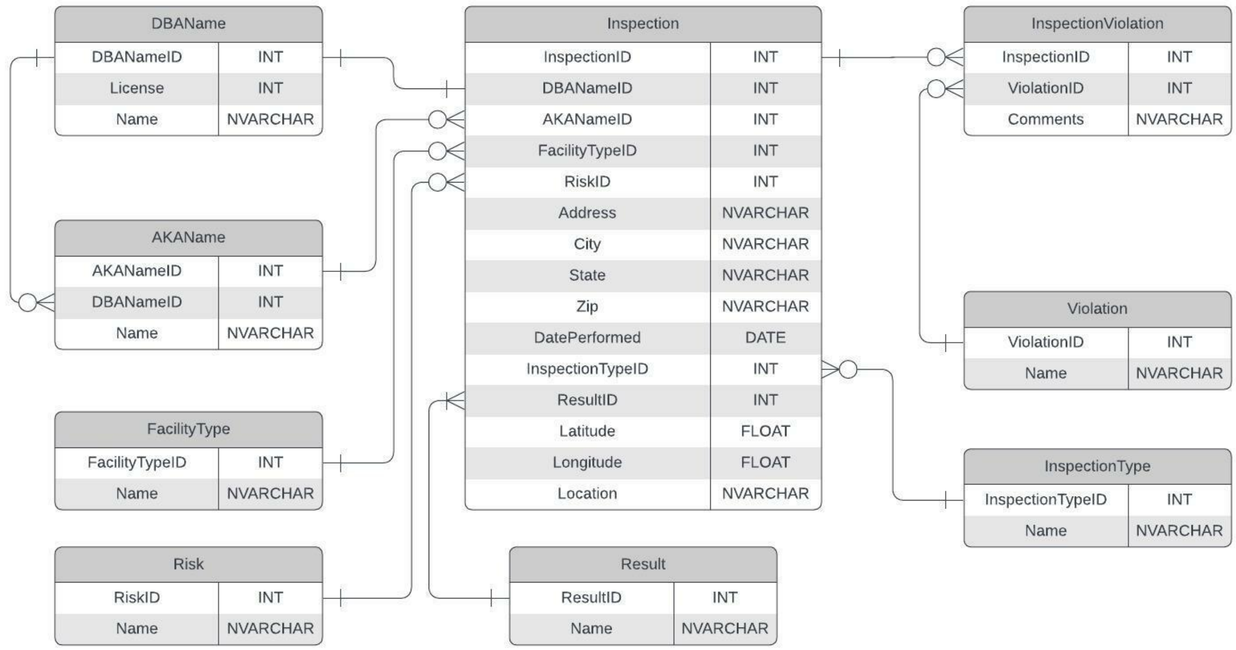
CSV Field	Schema Name	Description
Inspection ID	Inspection.InspectionID	A unique integer ID identifying the inspection.
DBA Name	DBAName.DBAName	"Doing business as". The legal name of the establishment.
AKA Name	AKAName.AKAName	"Also known as". The known name of the establishment to the public.
License #	DBAName.License	Unique license number assigned to the establishment by the Department of Business Affairs and Consumer Protection.
Facility Type	FacilityType.Name	A set of categories for the business' function (e.g., "Restaurant", "Grocery Store", etc.) of which one is assigned to the inspection. According to the official dataset documentation, only the following categories should be present: bakery, banquet

[2] https://data.cityofchicago.org/api/assets/BAD5301B-681A-4202-9D25-51B2CAE672FF

CSV Field	Schema Name	Description
		hall, candy store, caterer, coffee shop, day care center (ages less than 2), day care center (ages 2 – 6), day care center (combo, ages less than 2 and 2 – 6 combined), gas station, Golden Diner, grocery store, hospital, long term care center (nursing home), liquor store, mobile food dispenser, restaurant, paleteria, school, shelter, tavern, social club, wholesaler, or Wrigley Field Rooftop.
Risk	Risk.Name	Risk classification of the establishment, ranging from 1 to 3. The values are alphanumeric in the CSV file: "Risk 1 (High)", "Risk 2 (Medium)", and "Risk 3 (Low)".
Address	Inspection.Address	The street address of the business.
City	Inspection.City	The city in which the business is located.
State	Inspection.State	The state in which the business is located.
Zip	Inspection.Zip	The zip code in which the business is located.
Inspection Date	Inspection.DatePerformed	The date of inspection, format "MM/DD/YYYY".
Inspection Type	InspectionType.Name	Set of values categorizing the type of the inspection, of which one is assigned to the inspection. According to the documentation, the following types of inspections should be present: **canvass**: most common, perform at frequent intervals relative to the risk of the establishment. **consultation**: at the request of the owner prior to the opening of the establishment; **complaint**: inspection in response to a complaint against an establishment; **license**: inspection is done as a requirement for the establishment to receive its license to operate; **suspect food poisoning:** inspection is done in response to one or more claims of illness related to eating at the establishment; **task-force operation**: inspections of bars or taverns; **re-inspections**: can occur for most types of inspections.
Results	Result.Name	Set of values categorizing the outcome of the inspection of which one is assigned to the inspection. According to the documentation, the following results can be expected: pass, pass with conditions, fail, established not found or establishment out of business.

CSV Field	Schema Name	Description
Violations	Violation.Name	A large text string containing the concatenation of all of the violations in coded form which arose during the inspection as well as the inspector’s free text comments for each violation. According to the official documentation, there are 45 distinct violations numbered 1-44 and 70.
Latitude	Inspection.Latitude	The geographic latitude of the business.
Longitude	Inspection.Longitude	The geographic longitude of the business.
Location	Inspection.Location	An ordered pair of Latitude and Longitude.

These fields were normalized into a relational schema, displayed by the diagram below, followed by the SQL DDL for generation. It must be noted that no additional fields were introduced, however non-mapping tables were assigned surrogate integer primary keys.



```
CREATE TABLE FacilityType
(
       FacilityTypeID INT NOT NULL IDENTITY(1,1)
       ,[Name] NVARCHAR(100)

       ,CONSTRAINT PK_FacilityType PRIMARY KEY CLUSTERED (FacilityTypeID)
       ,CONSTRAINT AK_FacilityType_Name UNIQUE ([Name])
);
```

```
CREATE TABLE Risk
(
      RiskID INT NOT NULL IDENTITY(1,1)
      ,[Name] NVARCHAR(100)

      ,CONSTRAINT PK_Risk PRIMARY KEY CLUSTERED (RiskID)
      ,CONSTRAINT AK_Risk_Name UNIQUE ([Name])
);

CREATE TABLE InspectionType
(
      InspectionTypeID INT NOT NULL IDENTITY(1,1)
      ,[Name] NVARCHAR(100)

      ,CONSTRAINT PK_InspectionType PRIMARY KEY CLUSTERED (InspectionTypeID)
      ,CONSTRAINT AK_InspectionType_Name UNIQUE ([Name])
);

CREATE TABLE Result
(
      ResultID INT NOT NULL IDENTITY(1,1)
      ,[Name] NVARCHAR(100)

      ,CONSTRAINT PK_Result PRIMARY KEY CLUSTERED (ResultID)
      ,CONSTRAINT AK_Result_Name UNIQUE ([Name])
);

CREATE TABLE Violation
(
      ViolationID INT NOT NULL
      ,[Name] NVARCHAR(500) NOT NULL

      ,CONSTRAINT PK_Violation PRIMARY KEY CLUSTERED (ViolationID)
      ,CONSTRAINT AK_Violation_Name UNIQUE ([Name])
);

CREATE TABLE DBAName
(
      DBANameID INT NOT NULL IDENTITY(1,1)
      ,License INT NOT NULL
      ,[Name] NVARCHAR(500) NOT NULL

      ,CONSTRAINT PK_DBAName PRIMARY KEY CLUSTERED (DBANameID)
      ,CONSTRAINT AK_DBAName_License UNIQUE (License)
);

CREATE TABLE AKAName
(
      AKANameID INT NOT NULL IDENTITY(1,1)
      ,DBANameID INT NOT NULL
      ,[Name] NVARCHAR(500) NOT NULL
      ,CONSTRAINT PK_AKAName PRIMARY KEY CLUSTERED (AKANameID)
);

CREATE TABLE Inspection
(
      InspectionID INT NOT NULL
      ,DBANameID INT NOT NULL
      ,AKANameID INT NOT NULL
      ,FacilityTypeID INT NULL
      ,RiskID INT NULL
      ,[Address] NVARCHAR(250) NULL
      ,City NVARCHAR(250) NULL
      ,[State] NCHAR(2) NOT NULL
      ,Zip NCHAR(2) NOT NULL
      ,[DatePerformed] DATE NOT NULL
      ,InspectionTypeID INT NULL
      ,ResultID INT NOT NULL
      ,Latitude FLOAT NULL
      ,Longitude FLOAT NULL
      ,[Location] NVARCHAR(100) NULL

      ,CONSTRAINT PK_Inspection PRIMARY KEY CLUSTERED (InspectionID)
      ,CONSTRAINT FK_Inspection_DBANameID__DBAName_DBANameID FOREIGN KEY
(DBANameID) REFERENCES DBAName(DBANameID)
      ,CONSTRAINT FK_Inspection_AKANameID__AKAName_AKANameID FOREIGN KEY
(AKANameID) REFERENCES AKAName(AKANameID)
      ,CONSTRAINT FK_Inspection_FacilityTypeID__FacilityType_FacilityTypeID
FOREIGN KEY (FacilityTypeID) REFERENCES FacilityType(FacilityTypeID)
      ,CONSTRAINT FK_Inspection_RiskID__Risk_RiskID FOREIGN KEY (RiskID)
REFERENCES Risk(RiskID)
      ,CONSTRAINT
FK_Inspection_InspectionTypeID__InspectionType_InspectionTypeID FOREIGN KEY
(InspectionTypeID) REFERENCES InspectionType(InspectionTypeID)
      ,CONSTRAINT FK_Inspection_ResultID__Result_ResultID FOREIGN KEY
(ResultID) REFERENCES Result(ResultID)
);

CREATE TABLE InspectionViolation
(
      InspectionID INT NOT NULL
      ,ViolationID INT NOT NULL
      ,Comments NVARCHAR(MAX) NULL

      ,CONSTRAINT PK_InspectionViolation PRIMARY KEY CLUSTERED (InspectionID,
ViolationID)
      ,CONSTRAINT FK_InspectionViolation_InspectionID__Inspection_InspectionID
FOREIGN KEY (InspectionID) REFERENCES Inspection(InspectionID)
      ,CONSTRAINT FK_InspectionViolation_ViolationID__Violation_ViolationID
FOREIGN KEY (ViolationID) REFERENCES Violation(ViolationID)
      ,INDEX IX_InspectionViolation_ViolationID NONCLUSTERED (ViolationID)
);
```

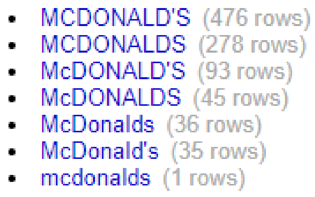
4. Dataset Quality Problems

In this section, data quality problems observed during initial analysis of the dataset are described in order to establish possible data cleaning measures necessary to support primary use case U_1.

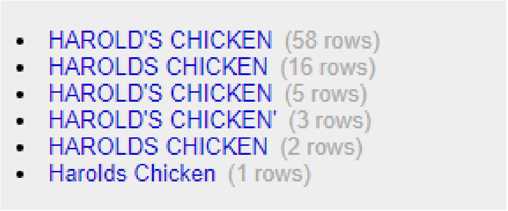
4.1 DBA Name

The **DBA Name** field contains text data entries with almost no consistency and various typographical errors. For example, 143,367 of 153,810 values (approximately 93%) are in uppercase, where the remaining values are in title case.

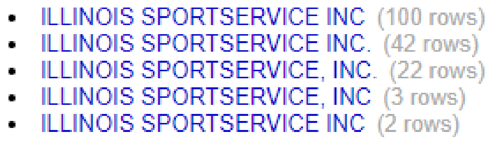
Multiple entries have several space characters, incorrect punctuation, and obvious (or suspected) misspellings. Furthermore, it was found that values representing the same business (identified by the business' license number) are spelled inconsistently across different records, as exemplified by the following clusters detected by OpenRefine:



- MCDONALD'S (476 rows)
- MCDONALDS (278 rows)
- McDONALD'S (93 rows)
- McDONALDS (45 rows)
- McDonalds (36 rows)
- McDonald's (35 rows)
- mcdonalds (1 rows)



- HAROLD'S CHICKEN (58 rows)
- HAROLDS CHICKEN (16 rows)
- HAROLD'S CHICKEN (5 rows)
- HAROLD'S CHICKEN' (3 rows)
- HAROLDS CHICKEN (2 rows)
- Harolds Chicken (1 rows)

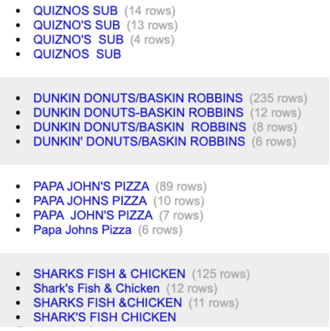


- ILLINOIS SPORTSERVICE INC (100 rows)
- ILLINOIS SPORTSERVICE INC. (42 rows)
- ILLINOIS SPORTSERVICE, INC. (22 rows)
- ILLINOIS SPORTSERVICE, INC (3 rows)
- ILLINOIS SPORTSERVICE INC (2 rows)

Therefore, this is a data quality problem that will need to be solved, since the business name should be normalized across all instances to allow the business name to be a search criteria per U_1. Without this previous normalization, a search may return partial or incorrect results based upon the varying spellings within the dataset.

4.2 AKA Name

The **AKA Name** field presents data quality issues similar to the **DBA Name** field, with OpenRefine detecting **598** possible clusters of business names, as exemplified by the left image below:

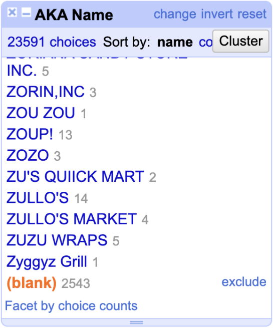


- QUIZNOS SUB (14 rows)
- QUIZNO'S SUB (13 rows)
- QUIZNO'S SUB (4 rows)
- QUIZNOS SUB

- DUNKIN DONUTS/BASKIN ROBBINS (235 rows)
- DUNKIN DONUTS-BASKIN ROBBINS (12 rows)
- DUNKIN DONUTS/BASKIN ROBBINS (8 rows)
- DUNKIN' DONUTS/BASKIN ROBBINS (6 rows)

- PAPA JOHN'S PIZZA (89 rows)
- PAPA JOHNS PIZZA (10 rows)
- PAPA JOHN'S PIZZA (7 rows)
- Papa Johns Pizza (6 rows)

- SHARKS FISH & CHICKEN (125 rows)
- Shark's Fish & Chicken (12 rows)
- SHARKS FISH &CHICKEN (11 rows)
- SHARK'S FISH CHICKEN



AKA Name change invert reset

23591 choices Sort by: name co Cluster

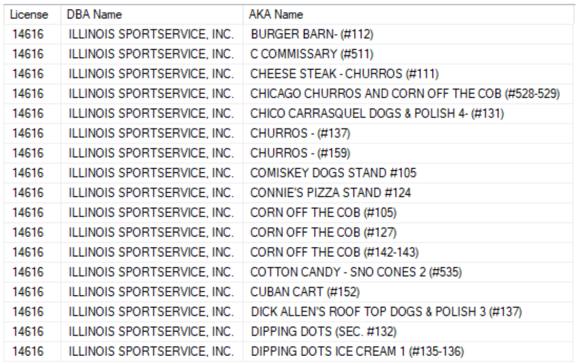
INC. 5
ZORIN,INC 3
ZOU ZOU 1
ZOUP! 13
ZOZO 3
ZU'S QUIICK MART 2
ZULLO'S 14
ZULLO'S MARKET 4
ZUZU WRAPS 5
Zyggyz Grill 1
(blank) 2543 exclude

Facet by choice counts

Further analysis indicates **2,543** records containing a **blank** value (right image above). However, this is not considered to be a data quality issue, as this specification describes this field as being a supplementary alias to the **DBA Name** field.

Additionally, the field appears to have a dual purpose since some large operations, such as sporting arenas, convention centers, and hotels, were issued a single (or a small number) of licenses, with multiple internal operations (such as concession stands) rolled up under a **single DBA Name**, using the **AKA Name** to distinguish between them.

For example, consider Illinois Sportsservice Inc., concessions operator at Guaranteed Rate Field, home of the Chicago White Sox baseball team. In total, their operational history contains **86** unique concessions operations, shown below as a subset:

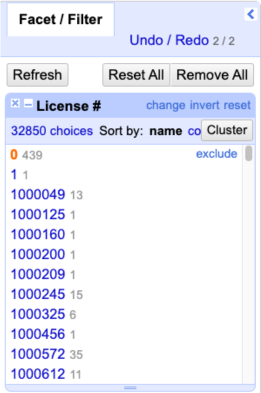


License	DBA Name	AKA Name
14616	ILLINOIS SPORTSERVICE, INC.	BURGER BARN- (#112)
14616	ILLINOIS SPORTSERVICE, INC.	C COMMISSARY (#511)
14616	ILLINOIS SPORTSERVICE, INC.	CHEESE STEAK - CHURROS (#111)
14616	ILLINOIS SPORTSERVICE, INC.	CHICAGO CHURROS AND CORN OFF THE COB (#528-529)
14616	ILLINOIS SPORTSERVICE, INC.	CHICO CARRASQUEL DOGS & POLISH 4- (#131)
14616	ILLINOIS SPORTSERVICE, INC.	CHURROS - (#137)
14616	ILLINOIS SPORTSERVICE, INC.	CHURROS - (#159)
14616	ILLINOIS SPORTSERVICE, INC.	COMISKEY DOGS STAND #105
14616	ILLINOIS SPORTSERVICE, INC.	CONNIE'S PIZZA STAND #124
14616	ILLINOIS SPORTSERVICE, INC.	CORN OFF THE COB (#105)
14616	ILLINOIS SPORTSERVICE, INC.	CORN OFF THE COB (#127)
14616	ILLINOIS SPORTSERVICE, INC.	CORN OFF THE COB (#142-143)
14616	ILLINOIS SPORTSERVICE, INC.	COTTON CANDY - SNO CONES 2 (#535)
14616	ILLINOIS SPORTSERVICE, INC.	CUBAN CART (#152)
14616	ILLINOIS SPORTSERVICE, INC.	DICK ALLEN'S ROOF TOP DOGS & POLISH 3 (#137)
14616	ILLINOIS SPORTSERVICE, INC.	DIPPING DOTS (SEC. #132)
14616	ILLINOIS SPORTSERVICE, INC.	DIPPING DOTS ICE CREAM 1 (#135-136)

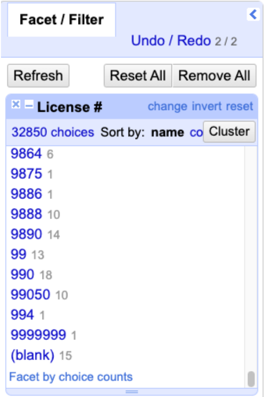
Given this scenario, the **AKA Name** field must be normalized so that it may be used as a search criteria alongside the **DBA Name** field, as per use case U_1. If this field isn't considered, a search for a “Dipping Dots” location at Guaranteed Rate Field would incorrectly return no results, as that business is rolled up the Illinois Sportservice Inc.

4.3 License

The official specification defines the **License #** field as containing the **unique** license number assigned for each business. However, there are **439** unique business records containing a license value of “0” and **15** records with an empty value, as demonstrated by the images below from OpenRefine:



0 439
1 1
1000049 13
1000125 1
1000160 1
1000200 1
1000209 1
1000245 15
1000325 6
1000456 1
1000572 35
1000612 11

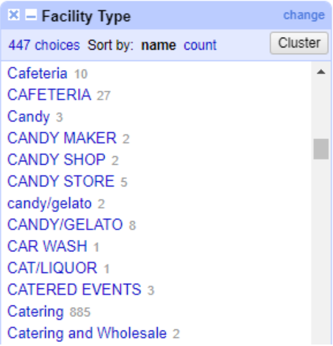


9864 6
9875 1
9886 1
9888 10
9890 14
99 13
990 18
99050 10
994 1
9999999 1
(blank) 15

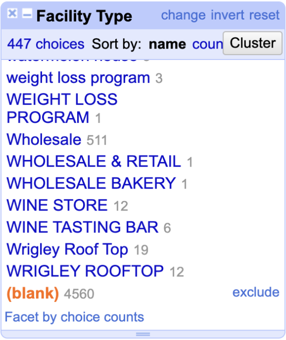
Hence, this subset of businesses represents an integrity constraint violation which will need to be corrected to satisfy U_1’s use of **License #** as a search criteria. This can be partially accomplished by attempting to match records with license number “0” to other records of the same business using the **DBA Name** and address fields.

4.4 Facility Type

The specification states that the **Facility Type** field categorizes the facility based upon a set of 23 types, as described in section 3. However, in dataset *D*, this field contains **447** unique values, including **4,560** records with no facility type indicated:

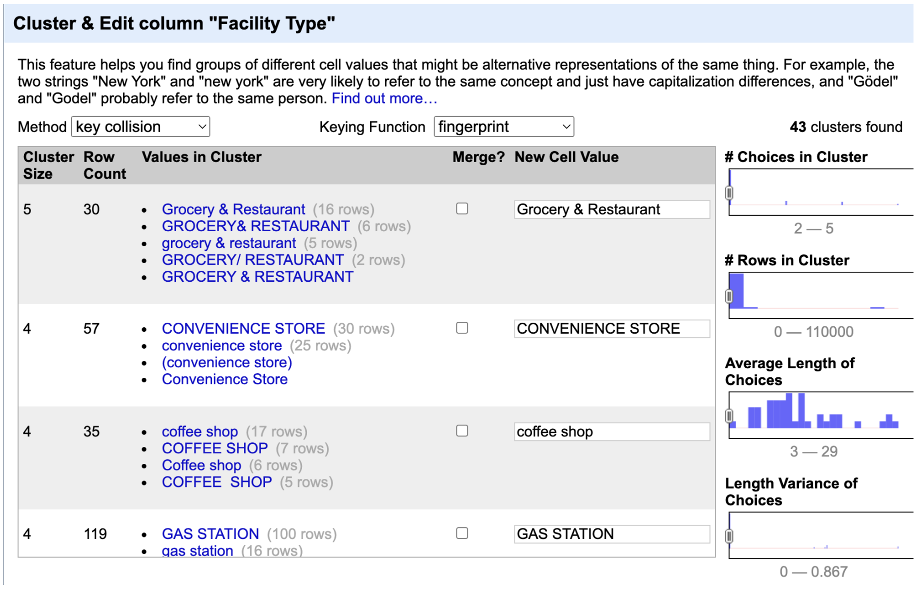


Cafeteria 10
CAFETERIA 27
Candy 3
CANDY MAKER 2
CANDY SHOP 2
CANDY STORE 5
candy/gelato 2
CANDY/GELATO 8
CAR WASH 1
CAT/LIQUOR 1
CATERED EVENTS 3
Catering 885
Catering and Wholesale 2



weight loss program 3
WEIGHT LOSS PROGRAM 1
Wholesale 511
WHOLESALE & RETAIL 1
WHOLESALE BAKERY 1
WINE STORE 12
WINE TASTING BAR 6
Wrigley Roof Top 19
WRIGLEY ROOFTOP 12
(blank) 4560

Furthermore, the field's values are inconsistent in formatting, have spelling errors, and contain duplication, as exemplified by the image below, indicating 43 possible clusters:



Cluster & Edit column "Facility Type"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

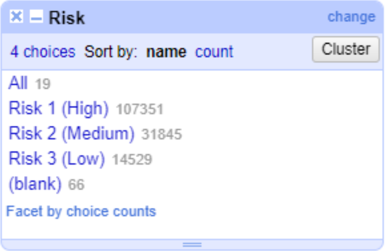
Method: key collision   Keying Function: fingerprint   **43** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	30	Grocery & Restaurant (16 rows) GROCERY& RESTAURANT (6 rows) grocery & restaurant (5 rows) GROCERY/ RESTAURANT (2 rows) GROCERY & RESTAURANT	□	Grocery & Restaurant
4	57	CONVENIENCE STORE (30 rows) convenience store (25 rows) (convenience store) Convenience Store	□	CONVENIENCE STORE
4	35	coffee shop (17 rows) COFFEE SHOP (7 rows) Coffee shop (6 rows) COFFEE SHOP (5 rows)	□	coffee shop
4	119	GAS STATION (100 rows) gas station (16 rows)	□	GAS STATION

In addition, some records have no obvious mapping back to the original values in the official documentation, for example, “CELL PHONE STORE”. These are likely errors at the point of data entry, and cannot be fully resolved to the specification from *D* alone.

Given these data quality problems, even though **Facility Type** is not used by U$_1$ as a search criteria, and, therefore, cleaning it is not necessary to make *D* fit-for-purpose, it may be partially done within our project to improve the quality of the query results.

4.5 Risk



Risk

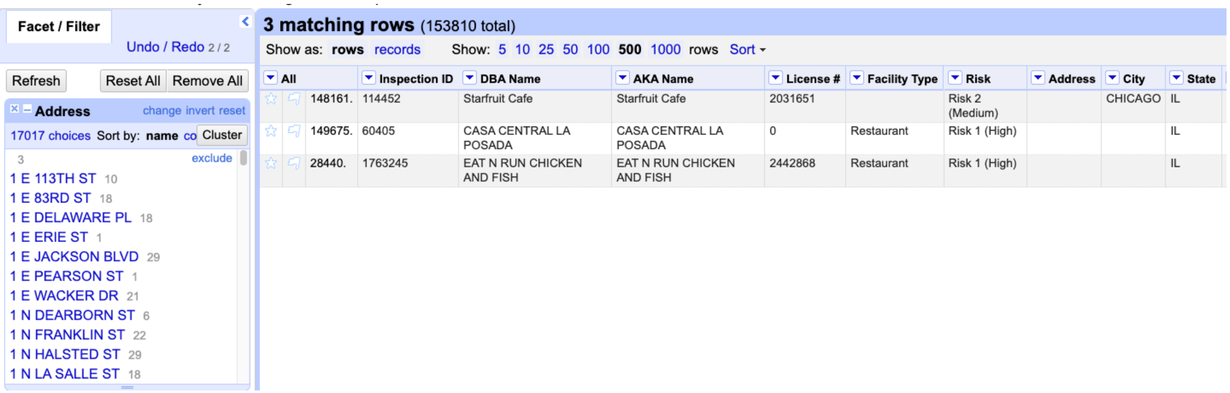
4 choices Sort by: name count

All	19
Risk 1 (High)	107351
Risk 2 (Medium)	31845
Risk 3 (Low)	14529
(blank)	66

The documentation indicates that the **Risk** field should be in the range of 1 (highest) to 3 (lowest). However, initial inspection of the dataset revealed **19** records with the value of "All" and **66** without a risk description, **0.01%** and **0.04%** of the dataset respectively, indicating a materially insignificant integrity constraint violation. Furthermore, since **Risk** is not directly used by U$_1$, cleaning it is not necessary to make *D* fit-for-purpose.

4.6 Address

As hinted by its name, the Address field contains the street address of the business and therefore should always be present. However, preliminary analysis indicated **3** records containing only spaces, with **2** of those not indicating the city. Nonetheless, they are materially insignificant, pertaining to only **0.002%** of the dataset.



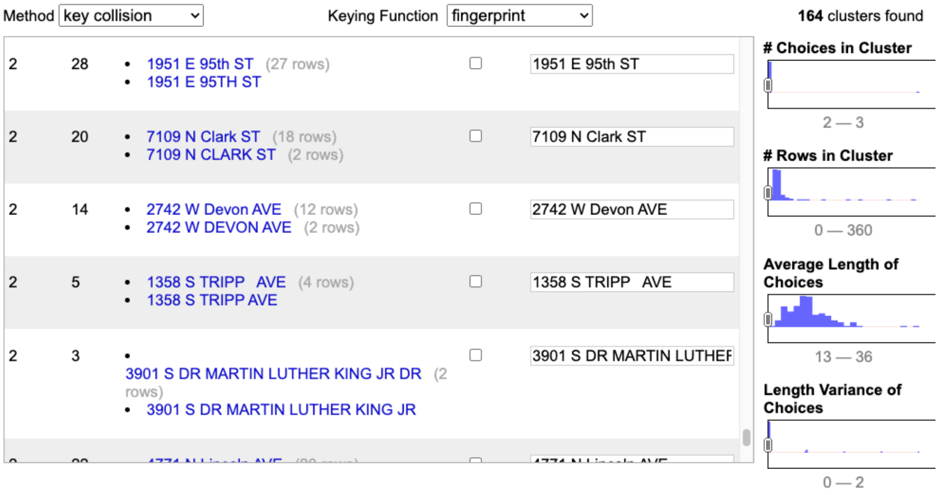
Address	
	3
1 E 113TH ST	10
1 E 83RD ST	18
1 E DELAWARE PL	18
1 E ERIE ST	1
1 E JACKSON BLVD	29
1 E PEARSON ST	1
1 E WACKER DR	21
1 N DEARBORN ST	6
1 N FRANKLIN ST	22
1 N HALSTED ST	29
1 N LA SALLE ST	18

3 matching rows (153810 total)

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

All	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State
148161.	114452	Starfruit Cafe	Starfruit Cafe	2031651		Risk 2 (Medium)		CHICAGO	IL
149675.	60405	CASA CENTRAL LA POSADA	CASA CENTRAL LA POSADA	0	Restaurant	Risk 1 (High)			IL
28440.	1763245	EAT N RUN CHICKEN AND FISH	EAT N RUN CHICKEN AND FISH	2442868	Restaurant	Risk 1 (High)			IL

In addition, OpenRefine detected **164** clusters as possibly the same addresses written in different forms. Therefore, the field, although not used by U_1 as a search parameter, may be partially cleaned in order to improve data quality of query results.

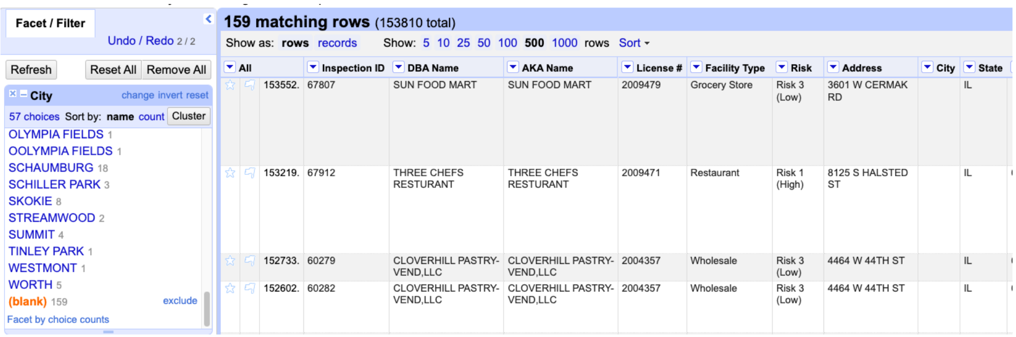


Method key collision Keying Function fingerprint **164** clusters found

2	28	1951 E 95th ST (27 rows) 1951 E 95TH ST	□	1951 E 95th ST
2	20	7109 N Clark ST (18 rows) 7109 N CLARK ST (2 rows)	□	7109 N Clark ST
2	14	2742 W Devon AVE (12 rows) 2742 W DEVON AVE (2 rows)	□	2742 W Devon AVE
2	5	1358 S TRIPP AVE (4 rows) 1358 S TRIPP AVE	□	1358 S TRIPP AVE
2	3	3901 S DR MARTIN LUTHER KING JR DR (2 rows) 3901 S DR MARTIN LUTHER KING JR	□	3901 S DR MARTIN LUTHEF

4.7 City

Regarding the **City** field, as can be seen in the table below, our initial analysis indicated that **159** records do not provide the city name (**blank**), roughly **0.10%** of the dataset.



Facet / Filter

Undo / Redo 2 / 2

Refresh | Reset All | Remove All

City — change invert reset

57 choices Sort by: name count Cluster

OLYMPIA FIELDS 1
OOLYMPIA FIELDS 1
SCHAUMBURG 18
SCHILLER PARK 3
SKOKIE 8
STREAMWOOD 2
SUMMIT 4
TINLEY PARK 1
WESTMONT 1
WORTH 5
(blank) 159 — exclude

Facet by choice counts

159 matching rows (153810 total)

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

All	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State
153552.	67807	SUN FOOD MART	SUN FOOD MART	2009479	Grocery Store	Risk 3 (Low)	3601 W CERMAK RD		IL
153219.	67912	THREE CHEFS RESTURANT	THREE CHEFS RESTURANT	2009471	Restaurant	Risk 1 (High)	8125 S HALSTED ST		IL
152733.	60279	CLOVERHILL PASTRY-VEND,LLC	CLOVERHILL PASTRY-VEND,LLC	2004357	Wholesale	Risk 3 (Low)	4464 W 44TH ST		IL
152602.	60282	CLOVERHILL PASTRY-VEND,LLC	CLOVERHILL PASTRY-VEND,LLC	2004357	Wholesale	Risk 3 (Low)	4464 W 44TH ST		IL

In addition, the name of the city Chicago is misspelled in various records, as exemplified in the table below. Furthermore, cities in the vicinity of Chicago are also present in the dataset, however this cannot be securely interpreted as a data quality issue.

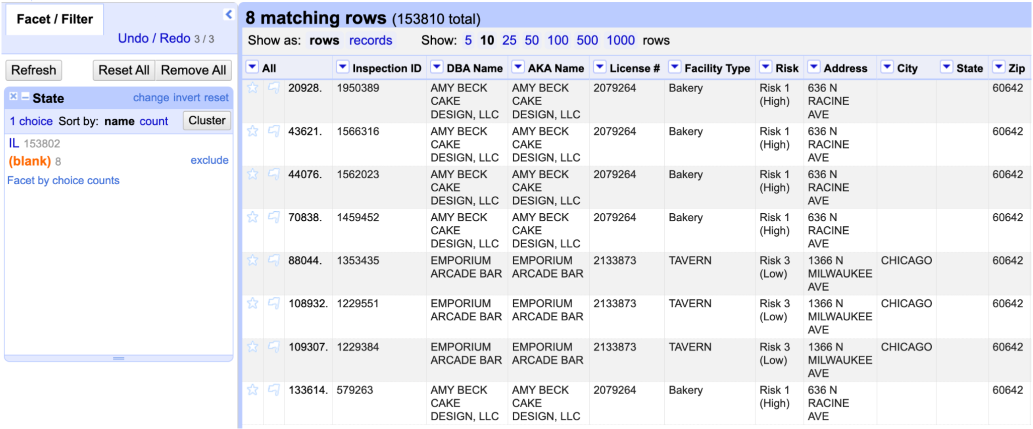
City	Qtd.	City	Qtd.	City	Qtd.
312CHICAGO	2	Chicago	258	MAYWOOD	14
ALSIP	3	CHicago	10	Maywood	1
alsip	1	CHICAGO HEIGHTS	2	NAPERVILLE	2
BANNOCKBURNDEERFIELD	2	CHICAGOCHICAGO	6	NILES NILES	3
BEDFORD PARK	2	CHICAGOI	3	Norridge	1
BERWYN	2	CICERO	6	OAK LAWN	1
BLOOMINGDALE	1	COUNTRY CLUB HILLS	1	OAK PARK	4
BLUE ISLAND	2	DES PLAINES	1	OLYMPIA FIELDS	1
BOLINGBROOK	1	EAST HAZEL CREST	3	OOLYMPIA FIELDS	1
BRIDEVIEW	1	ELK GROVE VILLAGE	12	SCHAUMBURG	18
BROADVIEW	1	ELMHURST	5	SCHILLER PARK	3
BURNHAM	1	EVANSTON	7	SKOKIE	8
CALUMET CITY	4	EVERGREEN PARK	1	STREAMWOOD	2
CCHICAGO	39	FRANKFORT	1	SUMMIT	4
CHARLES A HAYES	6	GLENCOE	1	TINLEY PARK	1
CHCHICAGO	6	INACTIVE	8	WESTMONT	1
CHCICAGO	3	JUSTICE	1	WORTH	5
CHESTNUT STREET	8	LAKE BLUFF	1	(blank)	159
CHICAGO	153090	LAKE ZURICH	1		
chicago	77	LOMBARD	1		

Given this scenario, although the **City** field is not utilized by U_1 as a search parameter, it may be partially cleaned in order to improve data quality of query results. Regarding the missing city names, they can be probably derived from the business address, since only **3** records in the dataset did not specify this last field.

4.8 State

Using OpenRefine, we detected that **8** records do not mention the State for the address. In addition, some of these records don't include a city name for the address, a problem that was detected in section 4.7.

However, it is known that all inspections occurred in the state of Illinois. A manual check of the other address fields of these 8 rows confirms this to be the case, as shown in the image below. Therefore, blank values can be corrected by converting them to “IL”.



State

1 choice Sort by: name count

IL 153802
(blank) 8

8 matching rows (153810 total)

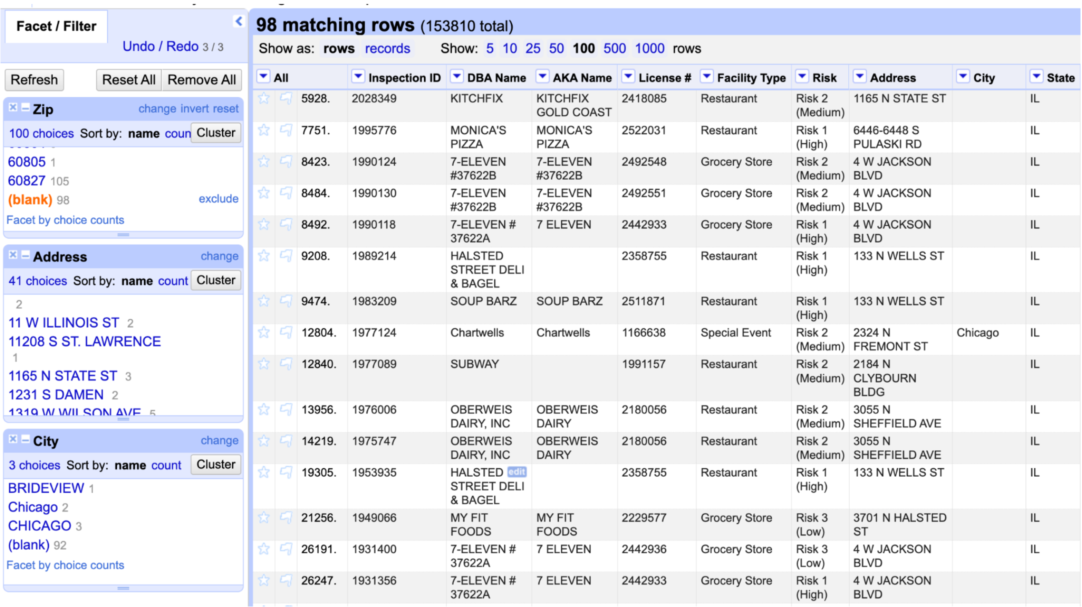
Show as: **rows** records Show: 5 **10** 25 50 100 500 1000 rows

All	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State	Zip
20928.	1950389	AMY BECK CAKE DESIGN, LLC	AMY BECK CAKE DESIGN, LLC	2079264	Bakery	Risk 1 (High)	636 N RACINE AVE			60642
43621.	1566316	AMY BECK CAKE DESIGN, LLC	AMY BECK CAKE DESIGN, LLC	2079264	Bakery	Risk 1 (High)	636 N RACINE AVE			60642
44076.	1562023	AMY BECK CAKE DESIGN, LLC	AMY BECK CAKE DESIGN, LLC	2079264	Bakery	Risk 1 (High)	636 N RACINE AVE			60642
70838.	1459452	AMY BECK CAKE DESIGN, LLC	AMY BECK CAKE DESIGN, LLC	2079264	Bakery	Risk 1 (High)	636 N RACINE AVE			60642
88044.	1353435	EMPORIUM ARCADE BAR	EMPORIUM ARCADE BAR	2133873	TAVERN	Risk 3 (Low)	1366 N MILWAUKEE AVE	CHICAGO		60642
108932.	1229551	EMPORIUM ARCADE BAR	EMPORIUM ARCADE BAR	2133873	TAVERN	Risk 3 (Low)	1366 N MILWAUKEE AVE	CHICAGO		60642
109307.	1229384	EMPORIUM ARCADE BAR	EMPORIUM ARCADE BAR	2133873	TAVERN	Risk 3 (Low)	1366 N MILWAUKEE AVE	CHICAGO		60642
133614.	579263	AMY BECK CAKE DESIGN, LLC	AMY BECK CAKE DESIGN, LLC	2079264	Bakery	Risk 1 (High)	636 N RACINE AVE			60642

With these observations, as noted for the **Address** and **City** fields, even though **State** is not used by U_1 as a search parameter, it will be cleaned to improve data quality.

4.9 Zip

In relation to the **Zip** field, our analysis detected **98** records not including a zip code for the business' address. A further look in OpenRefine revealed that **92** of these records also do not provide a city name, as exemplified below. However, these missing values can be derived from the business' address for **96** records.



Zip

100 choices Sort by: name coun

60805 1
60827 105
(blank) 98

Address

41 choices Sort by: name count

2
11 W ILLINOIS ST 2
11208 S ST. LAWRENCE 1
1165 N STATE ST 3
1231 S DAMEN 2
1319 W WILSON AVE 5

City

3 choices Sort by: name count

BRIDEVIEW 1
Chicago 2
CHICAGO 3
(blank) 92

98 matching rows (153810 total)

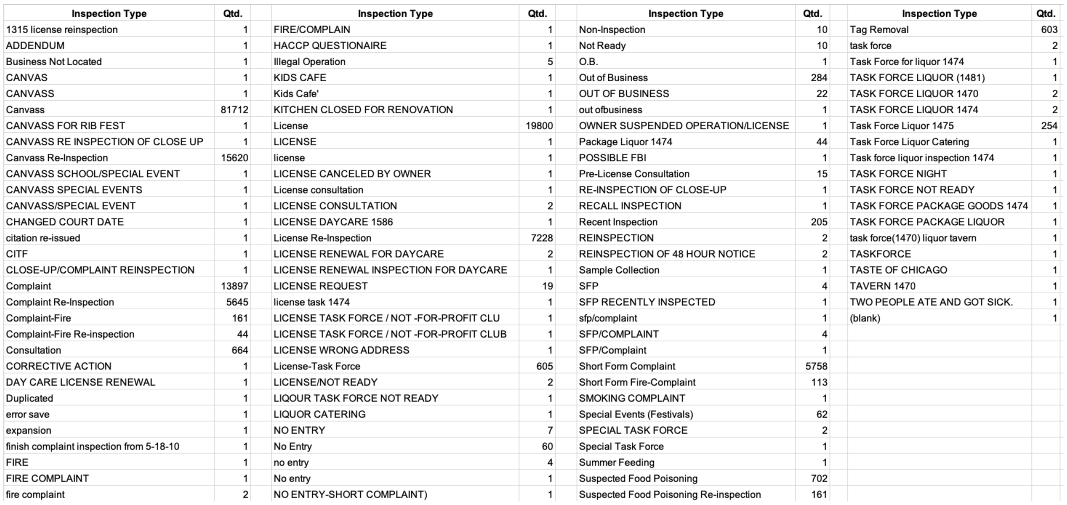
Show as: **rows** records Show: 5 10 25 50 **100** 500 1000 rows

All	Inspection ID	DBA Name	AKA Name	License #	Facility Type	Risk	Address	City	State
5928.	2028349	KITCHFIX	KITCHFIX GOLD COAST	2418085	Restaurant	Risk 2 (Medium)	1165 N STATE ST		IL
7751.	1995776	MONICA'S PIZZA	MONICA'S PIZZA	2522031	Restaurant	Risk 1 (High)	6446-6448 S PULASKI RD		IL
8423.	1990124	7-ELEVEN #37622B	7-ELEVEN #37622B	2492548	Grocery Store	Risk 2 (Medium)	4 W JACKSON BLVD		IL
8484.	1990130	7-ELEVEN #37622B	7-ELEVEN #37622B	2492551	Grocery Store	Risk 2 (Medium)	4 W JACKSON BLVD		IL
8492.	1990118	7-ELEVEN # 37622A	7 ELEVEN	2442933	Grocery Store	Risk 1 (High)	4 W JACKSON BLVD		IL
9208.	1989214	HALSTED STREET DELI & BAGEL		2358755	Restaurant	Risk 1 (High)	133 N WELLS ST		IL
9474.	1983209	SOUP BARZ	SOUP BARZ	2511871	Restaurant	Risk 1 (High)	133 N WELLS ST		IL
12804.	1977124	Chartwells	Chartwells	1166638	Special Event	Risk 2 (Medium)	2324 N FREMONT ST	Chicago	IL
12840.	1977089	SUBWAY		1991157	Restaurant	Risk 2 (Medium)	2184 N CLYBOURN BLDG		IL
13956.	1976006	OBERWEIS DAIRY, INC	OBERWEIS DAIRY	2180056	Restaurant	Risk 2 (Medium)	3055 N SHEFFIELD AVE		IL
14219.	1975747	OBERWEIS DAIRY, INC	OBERWEIS DAIRY	2180056	Restaurant	Risk 2 (Medium)	3055 N SHEFFIELD AVE		IL
19305.	1953935	HALSTED STREET DELI & BAGEL		2358755	Restaurant	Risk 1 (High)	133 N WELLS ST		IL
21256.	1949066	MY FIT FOODS	MY FIT FOODS	2229577	Grocery Store	Risk 3 (Low)	3701 N HALSTED ST		IL
26191.	1931400	7-ELEVEN # 37622A	7 ELEVEN	2442936	Grocery Store	Risk 3 (Low)	4 W JACKSON BLVD		IL
26247.	1931356	7-ELEVEN # 37622A	7 ELEVEN	2442933	Grocery Store	Risk 1 (High)	4 W JACKSON BLVD		IL

Therefore, as noted for the **Address**, **City and State** fields, even though **Zip** is not used by U_1 as a search parameter, it will be partially cleaned in order to improve data quality of query results.

4.10 Inspection Type

A brief look at contents of the **Inspection Type** field revealed that although the official specification presents **6** types of inspection (canvass, consultation, complaint, license, suspect food poisoning and task-force operation) and related re-inspections, the dataset contains **107** descriptions and **1** record with a **blank** value, as exemplified below:



Inspection Type	Qtd.	Inspection Type	Qtd.	Inspection Type	Qtd.	Inspection Type	Qtd.
1315 license reinspection	1	FIRE/COMPLAIN	1	Non-Inspection	10	Tag Removal	603
ADDENDUM	1	HACCP QUESTIONAIRE	1	Not Ready	10	task force	2
Business Not Located	1	Illegal Operation	5	O.B.	1	Task Force for liquor 1474	1
CANVAS	1	KIDS CAFE	1	Out of Business	284	TASK FORCE LIQUOR (1481)	1
CANVASS	1	Kids Cafe'	1	OUT OF BUSINESS	22	TASK FORCE LIQUOR 1470	2
Canvass	81712	KITCHEN CLOSED FOR RENOVATION	1	out ofbusiness	1	TASK FORCE LIQUOR 1474	2
CANVASS FOR RIB FEST	1	License	19800	OWNER SUSPENDED OPERATION/LICENSE	1	Task Force Liquor 1475	254
CANVASS RE INSPECTION OF CLOSE UP	1	LICENSE	1	Package Liquor 1474	44	Task Force Liquor Catering	1
Canvass Re-Inspection	15620	license	1	POSSIBLE FBI	1	Task force liquor inspection 1474	1
CANVASS SCHOOL/SPECIAL EVENT	1	LICENSE CANCELED BY OWNER	1	Pre-License Consultation	15	TASK FORCE NIGHT	1
CANVASS SPECIAL EVENTS	1	License consultation	1	RE-INSPECTION OF CLOSE-UP	1	TASK FORCE NOT READY	1
CANVASS/SPECIAL EVENT	1	LICENSE CONSULTATION	2	RECALL INSPECTION	1	TASK FORCE PACKAGE GOODS 1474	1
CHANGED COURT DATE	1	LICENSE DAYCARE 1586	1	Recent Inspection	205	TASK FORCE PACKAGE LIQUOR	1
citation re-issued	1	License Re-Inspection	7228	REINSPECTION	2	task force(1470) liquor tavern	1
CITF	1	LICENSE RENEWAL FOR DAYCARE	2	REINSPECTION OF 48 HOUR NOTICE	2	TASKFORCE	1
CLOSE-UP/COMPLAINT REINSPECTION	1	LICENSE RENEWAL INSPECTION FOR DAYCARE	1	Sample Collection	1	TASTE OF CHICAGO	1
Complaint	13897	LICENSE REQUEST	19	SFP	4	TAVERN 1470	1
Complaint Re-Inspection	5645	license task 1474	1	SFP RECENTLY INSPECTED	1	TWO PEOPLE ATE AND GOT SICK.	1
Complaint-Fire	161	LICENSE TASK FORCE / NOT -FOR-PROFIT CLU	1	sfp/complaint	1	(blank)	1
Complaint-Fire Re-inspection	44	LICENSE TASK FORCE / NOT -FOR-PROFIT CLUB	1	SFP/COMPLAINT	4		
Consultation	664	LICENSE WRONG ADDRESS	1	SFP/Complaint	1		
CORRECTIVE ACTION	1	License-Task Force	605	Short Form Complaint	5758		
DAY CARE LICENSE RENEWAL	1	LICENSE/NOT READY	2	Short Form Fire-Complaint	113		
Duplicated	1	LIQOUR TASK FORCE NOT READY	1	SMOKING COMPLAINT	1		
error save	1	LIQUOR CATERING	1	Special Events (Festivals)	62		
expansion	1	NO ENTRY	7	SPECIAL TASK FORCE	2		
finish complaint inspection from 5-18-10	1	No Entry	60	Special Task Force	1		
FIRE	1	no entry	4	Summer Feeding	1		
FIRE COMPLAINT	1	No entry	1	Suspected Food Poisoning	702		
fire complaint	2	NO ENTRY-SHORT COMPLAINT)	1	Suspected Food Poisoning Re-inspection	161		

Furthermore, for the documented inspection types, different descriptions were given, requiring data cleaning in order to group them. For example, canvass inspections are also referred to as "CANVAS", "CANVASS" or with other descriptors following its name such as "CANVASS SCHOOL/SPECIAL EVENT". The same can be seen for other types such as license, task-force, complaint, suspected food poisoning, this last one even referred in one record as "TWO PEOPLE ATE AND GOT SICK.".

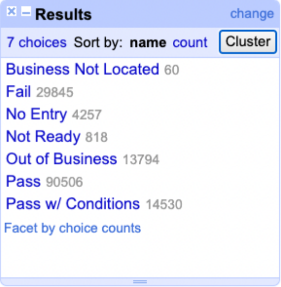
On the other hand, a significant number of inspections refer to certain types not present in the documentation, as exemplified by "Special Events (Festivals)" with **62** inspections and "Tag Removal" with **603** inspections. Nonetheless, these are materially insignificant, corresponding, respectively, to **0.04%** and **0.39%** of the dataset records.

In addition, it can be observed that some values don't seem to be related to the field, as exemplified by the following: "CHANGED COURT DATE", "CITF", "CORRECTIVE ACTION", "Duplicated", "error save", "expansion", "HACCP QUESTIONAIRE", "Illegal Operation", "KIDS CAFE", "KIds Cafe'", "Non-Inspection", "POSSIBLE FBI", "Sample Collection", "Summer Feeding", and others. Additionally, the following values apparently refer to inspection results instead of types: "NO ENTRY", "Not Ready", "OUT OF BUSINESS", "O.B", and "Business Not Located".

Given this scenario, although **Inspection Type** is not used by U_1, therefore cleaning it wouldn't be necessary to make *D* fit-for-purpose, it may be partially done to improve the quality of results and since it constitutes an integrity constraint violation according to the official documentation.

4.11 Results

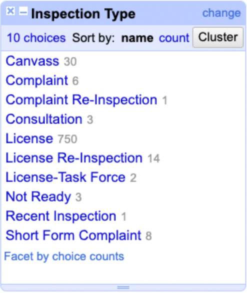
The **Results** field contains a description of the result of the inspection. According to the official specification, the following descriptors can be expected to exist in the dataset: pass, pass with conditions, fail, established not found or establishment out of business.



Results change
7 choices Sort by: name count Cluster

Business Not Located	60
Fail	29845
No Entry	4257
Not Ready	818
Out of Business	13794
Pass	90506
Pass w/ Conditions	14530

Facet by choice counts



Inspection Type change
10 choices Sort by: name count Cluster

Canvass	30
Complaint	6
Complaint Re-Inspection	1
Consultation	3
License	750
License Re-Inspection	14
License-Task Force	2
Not Ready	3
Recent Inspection	1
Short Form Complaint	8

Facet by choice counts

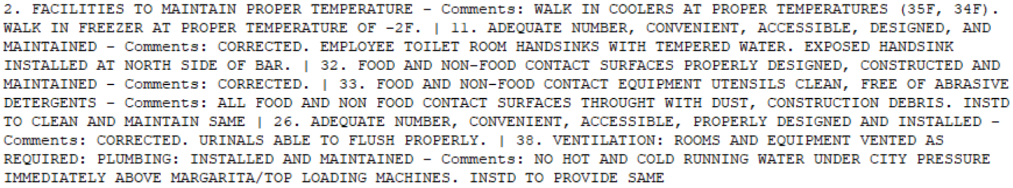
However, an initial look with OpenRefine revealed two additional descriptors within the dataset, although not indicated in the documentation: "No Entry" with **4,257** records and "Not Ready" with **818** records. A further inspection of the second group, indicates that it is mostly related to license inspections or reinspections (765 of 818 records), where the business is visited prior to receiving its license to operate. Therefore, it might indicate businesses that haven't finished renovations prior to the inspection or that were visited and found to not be ready in order to obtain a license.

Either way, the high number of records with these new descriptors indicates possible changes to the dataset not reflected in the documentation, which can cause integrity constraint violations if the last is considered as correct and adopted in our project.

Lastly, although this field is used by U_1 as a search parameter, this initial analysis does not indicate the need for data cleaning.

4.12 Violations

The **Violations** field contains a concatenated string of every violation encountered during a given inspection, as well as the inspectors' text comments for each violation (see the example figure). During our analysis, we found the string to be easily split by the pipe delimiter per the specification, and the comments easily split from the violation by using the comment prefix, “ - Comments:”, as a delimiter.

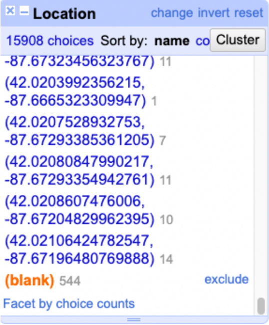


```
2. FACILITIES TO MAINTAIN PROPER TEMPERATURE - Comments: WALK IN COOLERS AT PROPER TEMPERATURES (35F, 34F).
WALK IN FREEZER AT PROPER TEMPERATURE OF -2F. | 11. ADEQUATE NUMBER, CONVENIENT, ACCESSIBLE, DESIGNED, AND
MAINTAINED - Comments: CORRECTED. EMPLOYEE TOILET ROOM HANDSINKS WITH TEMPERED WATER. EXPOSED HANDSINK
INSTALLED AT NORTH SIDE OF BAR. | 32. FOOD AND NON-FOOD CONTACT SURFACES PROPERLY DESIGNED, CONSTRUCTED AND
MAINTAINED - Comments: CORRECTED. | 33. FOOD AND NON-FOOD CONTACT EQUIPMENT UTENSILS CLEAN, FREE OF ABRASIVE
DETERGENTS - Comments: ALL FOOD AND NON FOOD CONTACT SURFACES THROUGHT WITH DUST, CONSTRUCTION DEBRIS. INSTD
TO CLEAN AND MAINTAIN SAME | 26. ADEQUATE NUMBER, CONVENIENT, ACCESSIBLE, PROPERLY DESIGNED AND INSTALLED -
Comments: CORRECTED. URINALS ABLE TO FLUSH PROPERLY. | 38. VENTILATION: ROOMS AND EQUIPMENT VENTED AS
REQUIRED: PLUMBING: INSTALLED AND MAINTAINED - Comments: NO HOT AND COLD RUNNING WATER UNDER CITY PRESSURE
IMMEDIATELY ABOVE MARGARITA/TOP LOADING MACHINES. INSTD TO PROVIDE SAME
```

The codes in the Violations field across the entirety of *D* are well-formed and conform to the specification, with the only data cleaning required being to split each inspection’s violation and comments (zero or more violations per inspection) from a single field into a mapping relation between inspections and the unique list of violations - a task which we already accomplished during initial analysis. After this normalization task is performed, instances of individual code violations can be easily queried as required by U_1.

4.13 Latitude, Longitude, and Location

Because of their relationship, we discuss the **Latitude**, **Longitude**, and **Location** fields in conjunction. The first two are the geographic latitude and longitude coordinate of the business, respectively, expressed in decimal degrees. The **Location** is approximately these values, within one-ten thousandth of a degree, expressed as an ordered pair.



15908 choices Sort by: name co

-87.67323456323767) 11
(42.0203992356215,
-87.6665323309947) 1
(42.0207528932753,
-87.67293385361205) 7
(42.02080847990217,
-87.67293354942761) 11
(42.0208607476006,
-87.67204829962395) 10
(42.02106424782547,
-87.67196480769888) 14
(blank) 544

Initial analysis in OpenRefine revealed that **544** records, representing **161** businesses, do not contain latitude, longitude, or location data, accounting for **0.35%** of the dataset.

Although U_1 proposes to make use of the geographical location to place locations on a graphical map, we still consider *D* fit-for-purpose given the insignificantly small number of records missing a location. In addition, the possibility remains that this incongruence can be partially reduced if we are able to recover the data from other records related to the same license number during the second phase of the project or derive the data from the available business address using tools such as Google Maps.

5. Project Plan

Having established our primary use case U_1, the structure of the dataset, its data quality problems and integrity constraints, our initial proposed project plan is as follows.

Activities	Member
1. Perform an analysis of the field-level data of the dataset Primary fields for U_1: Inspection ID, DBA Name, AKA Name, License #, Inspection Date, Results, Violations, Inspection Type. Secondary fields (partial cleaning to improve quality of results related to U_1): Address, City, State, Zip, Latitude, Longitude, Location, Facility Type. Tools: OpenRefine, SQL.	Steve Fabricio
2. Perform syntactic and semantic/integrity constraint corrections at column level related to U_1 Primary fields for U_1: Inspection ID, DBA Name, AKA Name, License #, Inspection Date, Results, Violations, Inspection Type. Secondary fields (partial cleaning to improve quality of results related to U_1): Address, City, State, Zip, Latitude, Longitude, Location, Facility Type. Tools: OpenRefine, SQL, Google Maps.	Steve Fabricio
3. Load output of OpenRefine into a staging table in SQL and correct schema-level integrity constraints Fields: License #, Results, Violations, Facility Type, Risk, Inspection Type. Tools: OpenRefine, Microsoft SQL Server, SQL.	Steve
4. Load the staging table into the integrity constraint-enforced SQL schema Fields: License #, Results, Violations, Facility Type, Risk, Inspection Type. Tools: Microsoft SQL Server, SQL.	Steve
5. Implement query(s) demonstrating resulting data set can successfully achieve use case U_1 Tools: Microsoft SQL Server, SQL.	Steve Fabricio
5.1 Integrate the resulting database with Tableau Public to create a dashboard visualization that exemplifies U_1 (optional, if time allows). Tools: Microsoft SQL Server, Tableau Desktop, Tableau Public.	Fabricio
6. Document changes to dataset and steps during data cleaning process (continuous process) Tools: Word, Excel, Draw.io, Git, OpenRefine recipes, Tableau project, text files, json files.	Team
7. Write phase 2 project report, develop illustrations and workflow diagram Tools: Word, Excel, Visio, Draw.io.	Roberto

References

[1] Chicago Restaurant Inspections. (2017, August 30). Kaggle. Retrieved June 22, 2022, from https://www.kaggle.com/datasets/chicago/chi-restaurant-inspections

[2] Chicago Data Portal. (n.d.). Food Inspections. Retrieved June 28, 2022, from https://data.cityofchicago.org/api/assets/BAD5301B-681A-4202-9D25-51B2CAE672FF