INFO399 Final Report

Jiayi Pan

[1. Use Case 1](#_Toc326)

[1.1 Dataset Description 1](#_Toc6976)

[1.2 Use Case 1](#_Toc20259)

[1.3 Targeted Columns 1](#_Toc23537)

[1.4 Tool Used 2](#_Toc29264)

[2. Dirty Data Analysis 2](#_Toc5706)

[1.1 Use Case 1 2](#_Toc23753)

[1.1.1 Zip vs. Results 2](#_Toc31477)

[1.1.2 Facility Type vs. Results 3](#_Toc28841)

[1.2 Use Case 2 6](#_Toc28498)

[1.2.1 Results of Facility Type vs. Inspection Date 6](#_Toc8890)

[1.2.2 Risk of Facility Type vs. Inspection Date 7](#_Toc20040)

[1.2.3 DBA Name vs. AKA Name 8](#_Toc16004)

[3. Initial Data Cleaning 9](#_Toc2188)

[3.1 Deal with Use Case 1 9](#_Toc16362)

[3.1.1 Step 1: Basic Cleaning 9](#_Toc26717)

[3.1.2 Step 2: Handle Facility Type 11](#_Toc5778)

[3.1.3 Step 3: Handle Results 13](#_Toc7096)

[3.1.4 Step 4: Handle Zip 15](#_Toc2475)

[3.2 Deal with Use Case 2 17](#_Toc401)

[3.2.1 Step 1: Handle Inspection Date 17](#_Toc26322)

[3.2.2 Step 2: Handle AKA Name 19](#_Toc22947)

[4. Evaluation 22](#_Toc15649)

[5. Refined Data Cleaning 23](#_Toc13254)

[5.1 Basic Cleaning 23](#_Toc11691)

[5.2 Facility Type Refined Cleaning 24](#_Toc8944)

[5.2 Other Refined Cleaning 31](#_Toc25846)

[6. Data Visualization and Evaluation 32](#_Toc6478)

[6.1 Pass Rate of Facility Type 32](#_Toc11576)

[6.2 Pass Rate at Each Region (Zip code) 35](#_Toc15896)

[6.3 Pass Rate of McDonald’s & Subway over Time 35](#_Toc18256)

[6.4 Risk Rate of Facility Type 37](#_Toc21260)

[7. Conclusion 38](#_Toc26851)

[Appendix 39](#_Toc11952)

[1. Summary Table 39](#_Toc28672)

[1.1 Initial cleaning summary 39](#_Toc14805)

[1.2 Defined cleaning summary 40](#_Toc8501)

[2. Overall Workflow 42](#_Toc4587)

# Use Case

## 1.1 Dataset Description

Restaurant inspections are conducted to ensure that licensed food establishments follow food safety guidelines when serving food to the public. The Food Protection Division of the Chicago Department of Public Health (CDPH) is dedicated to upholding the safety of food sold, bought, or prepared for public consumption in Chicago by conducting science-based inspections of all retail food establishments. All restaurants are required to undergo periodic inspections annually to ensure compliance with city ordinances and regulations. Accredited sanitarians licensed by CDPH conduct inspections on a variety of retail food suppliers including restaurants, grocery stores, bakeries, convenience stores, hospitals, nursing homes, day care facilities, shelters, schools, and temporary food service events.

The dataset has been compiled into a file called "Food\_Inspections.csv", which contains information on inspection dates, results, violations recorded, business names and their corresponding latitude/longitude coordinates, license numbers, and risk levels. The data covers the time period: 01/02/2010-08/28/2017.

## 1.2 Use Case

1. What is the pass rate of different food supplier types in Facility Type? What conclusions can be drawn? Which zip codes have the highest inspection passing rate? What is the geographical distribution of pass rates in different regions? (>=30)
2. From 2010 to 2017, what are the changes in inspection results for overall facility types? How does the overall risk level change? How does the inspection results for restaurants change? How does the inspection results for McDonald’s change?

## 1.3 Targeted Columns

• DBA Name (Doing Business As Name): the name that a company does business under, also known as a trade name, brand name, or business name.

• AKA Name (Also Known As Name): the alternative names that a company may use.

• Facility Type: the type or category of a food supplier, such as restaurants, grocery stores, bakeries, convenience stores, hospitals, nursing homes, day care facilities, shelters, schools, and temporary food service events.

• 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. (3 categories: Risk 1 (High), Risk 2 (Median), Risk 3 (Low)).

• Zip: a series of digits representing a specific geographical area of a food supplier

• Inspection Date: the specific date on which a food safety inspection is conducted at a particular food supplier (period: 01/02/2013-08/28/2017).

• Results: the outcome of a food safety inspection (7 categories: Business Not Located, Fail, No Entry, Not Ready, Out of Business, Pass, Pass w/ Conditions).

## 1.4 Tool Used

OpenRefine, Tableau, R

# Dirty Data Analysis

Before cleaning the raw data, it is necessary to perform data visualization and statistical analysis on the raw data to identify the possible issues that may serve as targets and directions for data cleaning.

## 1.1 Use Case 1

Data visualization and analysis on several key components from raw data (Zip, Facility Type, and Results) should be conducted, which can be used to inspire how to resolve underlying issues in use case a).

### 1.1.1 Zip vs. Results

The histogram of Zip vs. Results can be obtained by using Tableau, which focus on the distribution of the frequency of Zip codes based on 7 different Results.

Figure 1. Zip vs. Results

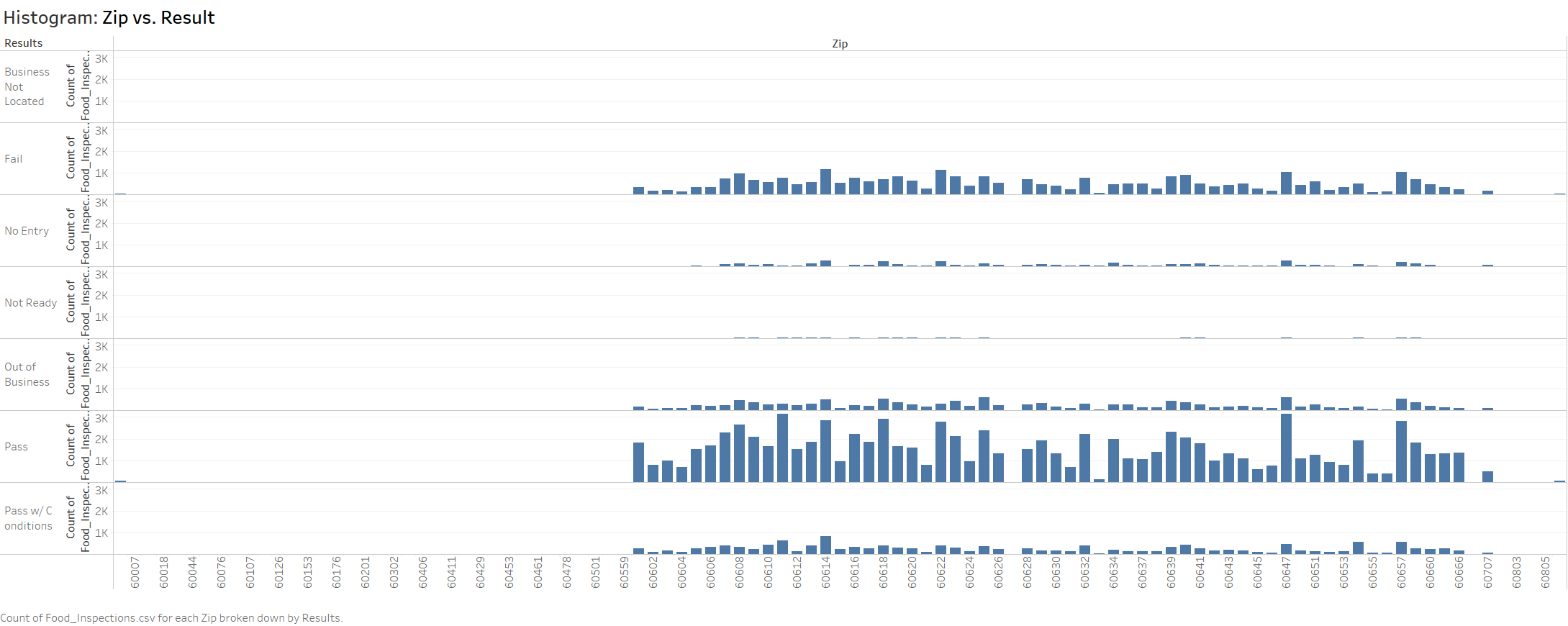
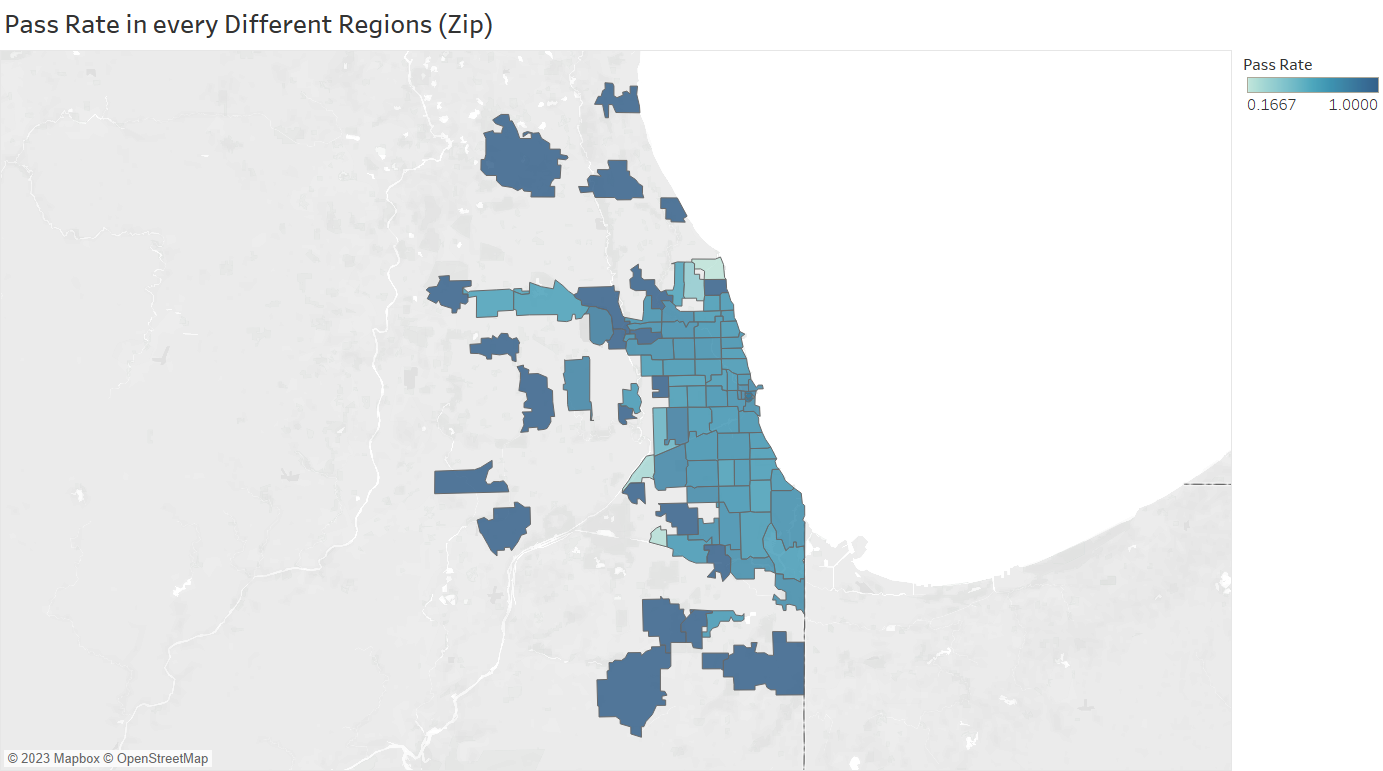


Figure 2. Pass Rate in every Different Regions (Zip)



The Figure#1 and Figure#2 reveal some issues that must be addressed in the initial data cleaning process:

1. The range of recorded Zip codes is from 60007 to 60850, but only the Zip code range from 60601 to 60707 is statistically significant with enough records, and therefore must be taken into consideration during the initial data cleaning process. To simplify the subsequent data processing, all Zip codes and their corresponding results that fall outside this significant range should be discarded.
2. The frequency of results corresponding to Zip codes has some statistical significance, but it is strongly correlated with the total number of food suppliers in a particular area. Therefore, a better way to handle the data is to investigate the passing rates in different Zip codes (areas). The Zip code range for processing is from 60601 to 60707, and some null values, extreme values or statistically insignificant data in this range need to be manually cleaned.
3. Tableau can recognize and convert Zip Codes into geographical coordinates, enabling the visualization on a map. Thus, it is meaningful to create a geographical distribution map based on Zip Codes and their corresponding area passing rates, which can be analyzed to explore the potential relationship between passing rates and geographical location. However, in Figure#2, it is noticed that the pass rate ranges from 0.1667 to 1.0. Based on the observation in Figure#1, it implies that there exist some outliers that contributes to this situation. Especially when the observation is too small, this issue is more likely to happen.

### 1.1.2 Facility Type vs. Results

The histogram of Facility Type vs. Frequency of Results focuses on the distribution of the Top 10 frequency of facility types based on 7 different Results.

Figure 3. Facility Type vs. Frequency of Results (Top10)

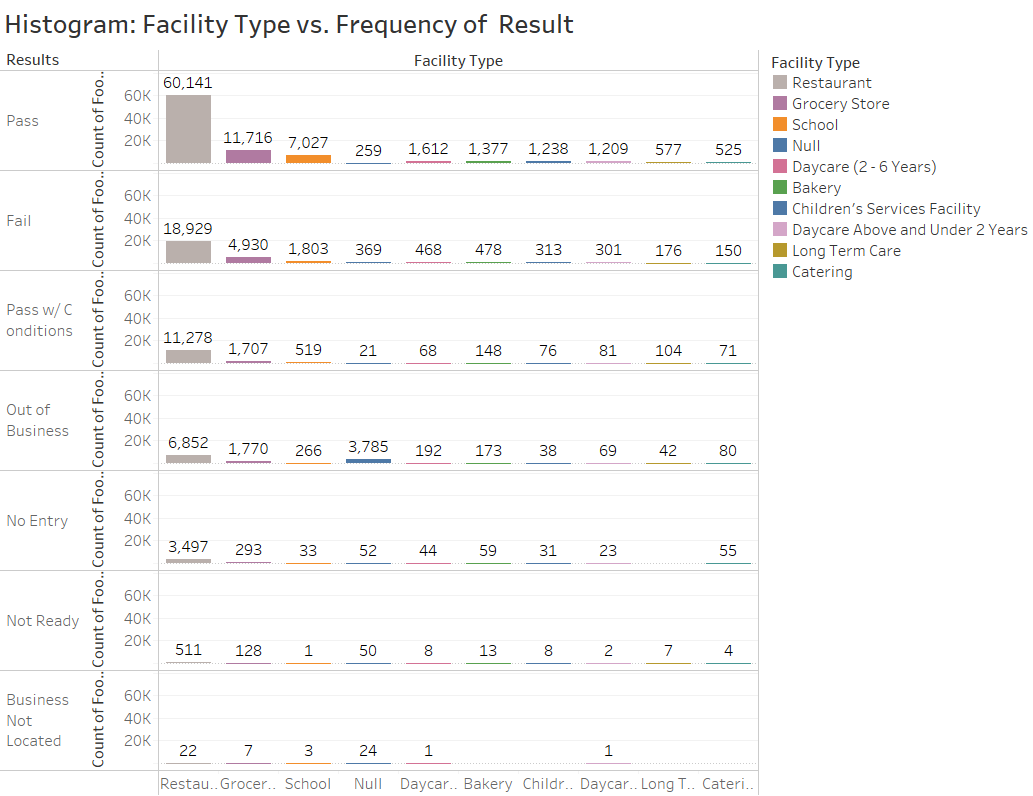
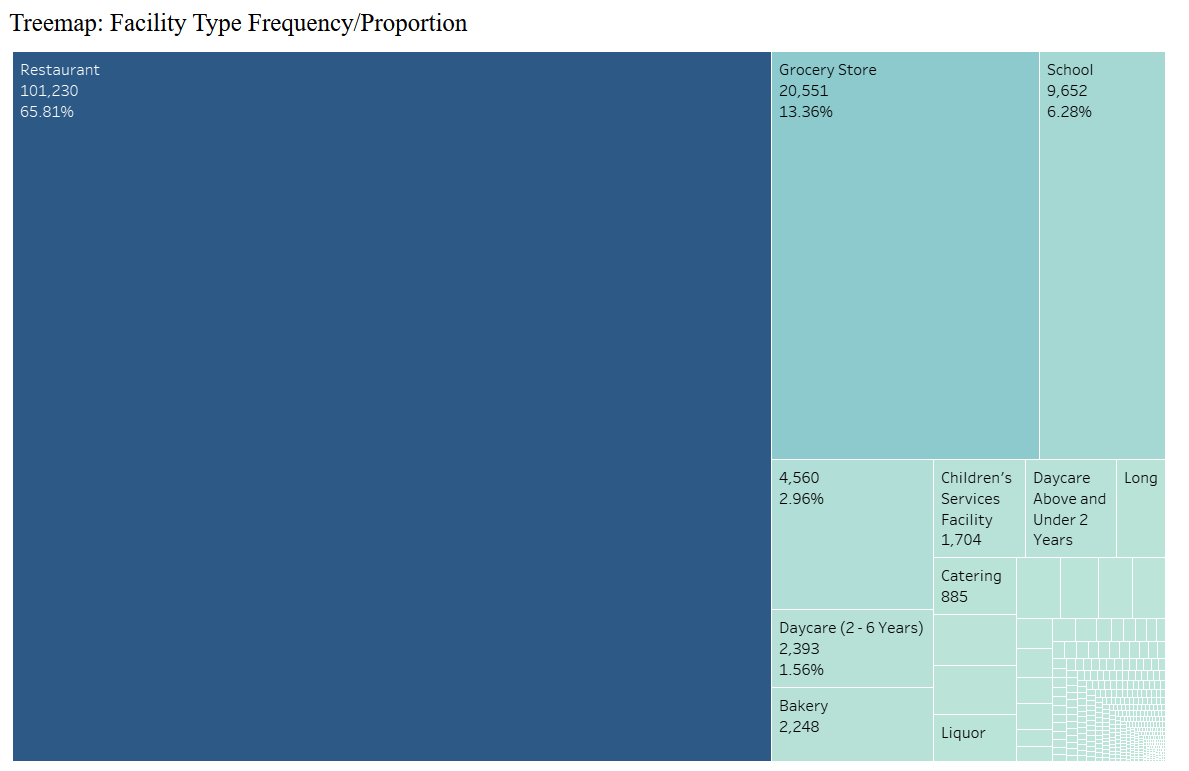


Figure 4. Treemap: Facility Type Frequency/Proportion



The histogram (graph#3) and treemap (graph#4) reveals some issues that must be addressed in the initial data cleaning process:

1. There are too many detailed facility types. It contributes to a rapid decrease in statistical significance of other facility types compared to one of the facility type: Restaurants, which is always with the highest frequency among Results categories. This makes it difficult to obtain a meaningful comparison of frequency for the other facility types. Therefore, it is necessary to merge and shrink facility types. Identical and similar types should be combined to finally obtain a smaller number of comparable major facility types.
2. In Figure#4, it is evident that the frequency of Facility Type: Restaurant is the highest. Its frequency is up to 101230, which accounts for 65.81% of the total, while the frequencies of other facility types significantly decrease and far lesser. It is clear that a large proportion of the investigated food suppliers are restaurants, which means that comparing the inspection results based on frequency alone is ineffective. Therefore, it is more important to refine the facility types and focus on the passing rate rather than passing frequency when analyzing the data.
3. In Figure#4, the frequency of nulls is 4560, which also take up 2.96% in the whole facility data. Thus, these null values should be removed to make sure the significance levels of meaningful facility types can be enhanced.

When the frequencies of Results for each Facility Type are ignored and paying attention to get the pass rate from the raw data, then the view version is shown in Table#1:

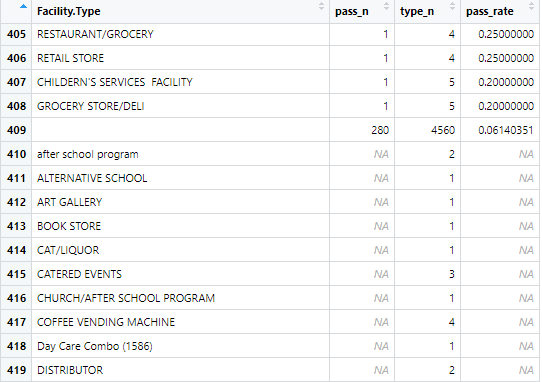
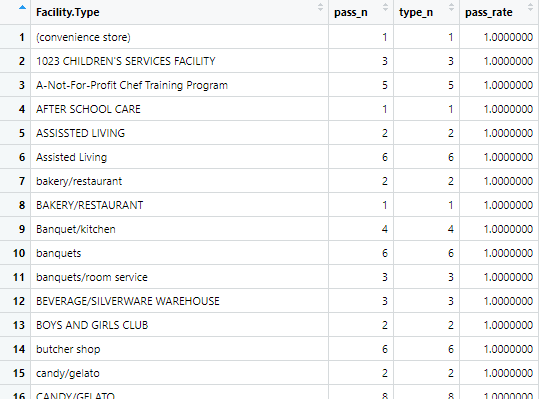


Table 1. Pass Rate of Facility Type

The final outlook consists of 448 rows, many of which show a pass rate of 1 for Facility Types due to a small number of surveys conducted. Additionally, there are some Facility Types with a pass rate displayed as NA because there are null values in their corresponding Results. Furthermore, there are observations with null Facility Types, which are separated into a separate row for observation.

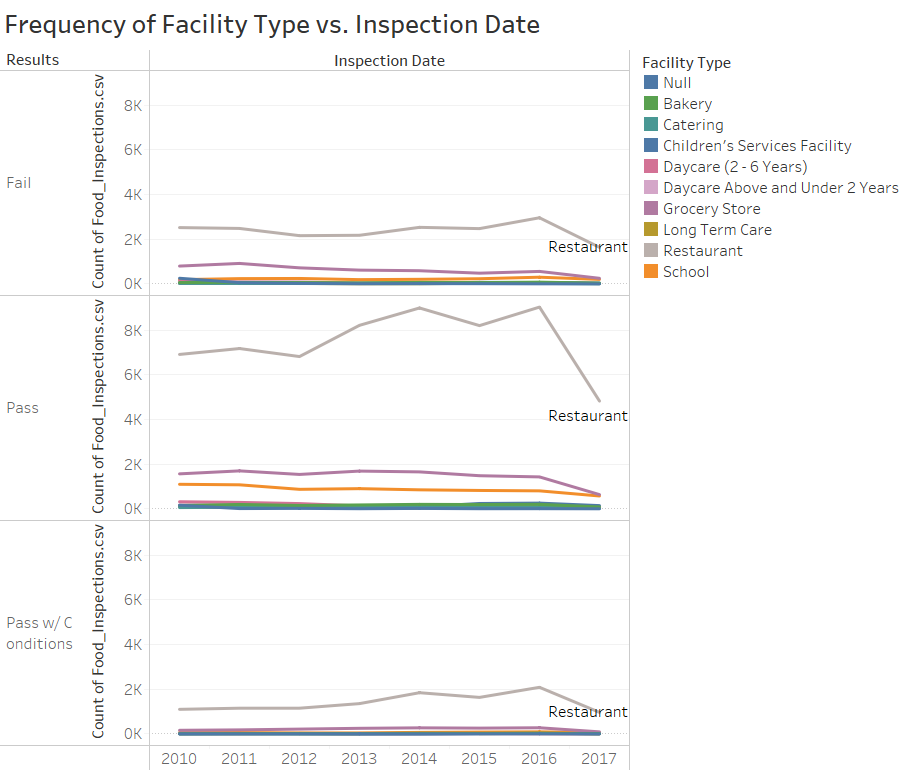
## 1.2 Use Case 2

In this part, data visualization and statistics analysis should be done on several crucial elements (DBA/AKA Name, Facility Type, Inspection Date, Risk) related to use case 2, in order to detect potential problems in advance and prepare for future initial data cleaning.

### 1.2.1 Results of Facility Type vs. Inspection Date

The line chart of Frequency of Facility Type vs. Inspection Date illustrates the trend of inspection results for food suppliers over time, with food inspections conducted each year.

Figure 5. Frequency of Facility Type vs. Inspection Date (Top10)



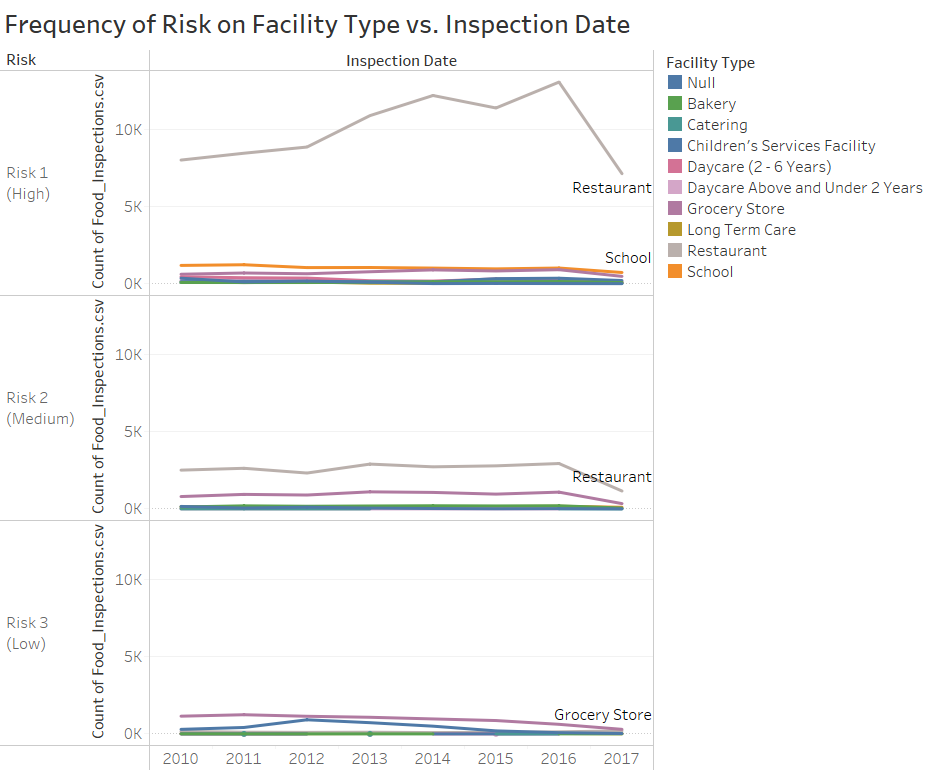
The line Figure#5 indicates several issues regarding inspection date and results situation of different facility types:

1. The frequency of facility types varies significantly depending on the type itself. Therefore, it is necessary to further reduce the types of facilities to simplify data analysis. Besides, it also seems that the annual frequency of passed food supplier is meaningless and it is better to use annual passing rate to make decisions.
2. All results show a downward trend in 2017. This is because the inspection date covers data from January 2010 to August 2017, so the data for the entire year of 2017 is not included in the inspection situation. Therefore, when exploring the trend of facility passing rate over the years in the future, the data from 2017 needs to be discarded. However, some data from 2017 can be used to help observe whether the predicted 2017 facility passing rate is accurate.

### 1.2.2 Risk of Facility Type vs. Inspection Date

The line chart of Frequency of Risk on Facility Type vs. Inspection Date illustrates the trend of inspected passing risk on food suppliers over time, based on food annual inspections.

Figure 6. Frequency of Risk on Facility Type vs. Inspection Date (Top10)



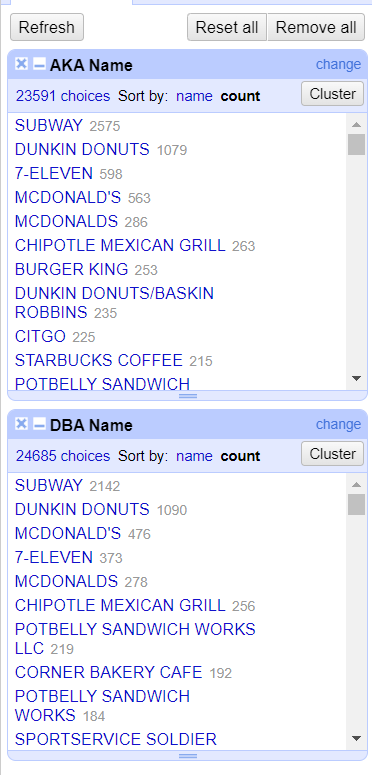
The line Figure#6 reveals some issues about inspection date and risk situation of different facility types:

1. In addition to the three main Risk levels (Risk 1, 2, and 3), there are also a small number of “Null” values and "All". To make the data more meaningful for further analysis, these two small parts of the data can be discarded.
2. The reason for the large difference among the three risk groups is due to the large difference in the of Facility Types and the imbalance in Risk assessment. It can be seen that due to strict inspections, most food suppliers' risk ratings fall into the Risk1: High category. Therefore, it would be more meaningful to compare the risk evaluation of facility types by introducing the risk rate rather than simply looking at the frequency of risk.
3. Since the data for 2017 does not cover the full year of risk statistics, it can be excluded and used as a test sample for comparison with estimated 2017 risk rates.

### 1.2.3 DBA Name vs. AKA Name

The food suppliers’ names listed in DBA Name and AKA Name can be compared by using Text Facet in OpenRefine directly.

Figure 7. DBA Name vs. AKA Name



Theoretically, the food suppliers’ name recorded under the DBA Name and AKA Name should be the same. However, in Figure#7, it can be observed that the total number of company names that include "food supplier" in the AKA Name column is 23,591, which is lower than the total listed in the DBA Name column (24,685). This suggests that AKA Name encompasses a broader and more accurate range of food suppliers. As a result, selecting AKA Name when dealing with a particular food supplier's name may potentially save time and costs associated with data cleaning, and may also yield more accurate results.

# Initial Data Cleaning

## Deal with Use Case 1

### 3.1.1 Step 1: Basic Cleaning

Firstly, some basic data cleaning processes need to be implemented. For example, in this section, the target columns should be filtered and corresponding null values need to be deleted.

Firstly, the non-target columns that are not related to the use cases need to be removed, and the target columns need to be retained (AKA Name, Facility Type, Zip, Inspection Date, Resultss, Risk). Afterwards, these six target columns can be reorganized to facilitate the data cleaning process and make it easier for data analysis in the subsequent steps (Table#2).

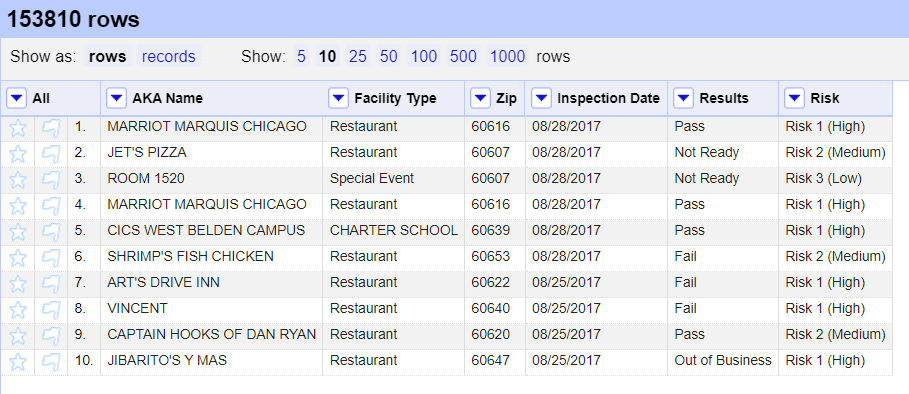


Table 2. Cleaned version 1

The next step is to handle null values. Since there are no numeric variables, the simplest method to handle null values can be used, which is to remove the null values in each column.

As shown in table#3, all the null values occurred in each columns are removed. The consequence is that the total rows drop from 151267 to 146776.

In fact, it is not a proper way to simply eliminate all null values as they may contain a lot of valuable information.

One approach to reducing the deletion of null values is to only search for and remove them in a small number of relevant columns. For instance, when addressing the issue of facility pass rate geographic distribution, it is actually only necessary to process the Zip, Results, and Facility Type columns. Therefore, to preserve useful information to the greatest extent possible, it is sufficient to delete null values only in these three columns rather than across the entire dataset.

Another approach is to fill null values as much as possible with external information. For example, DBA Name and AKA Name can be used to reference and fill each other. Also, Facility Type can be determined by referencing the DBA/AKA name of the food supplier. If the name contains sufficient information, it can more accurately determine the location of the food supplier and fill in its corresponding Zip Code.

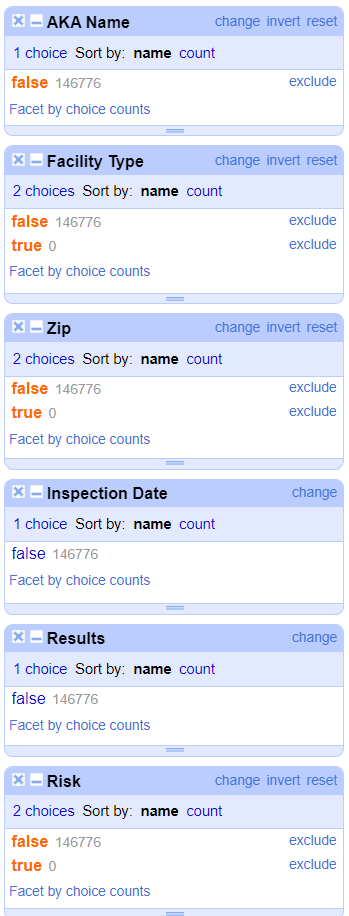
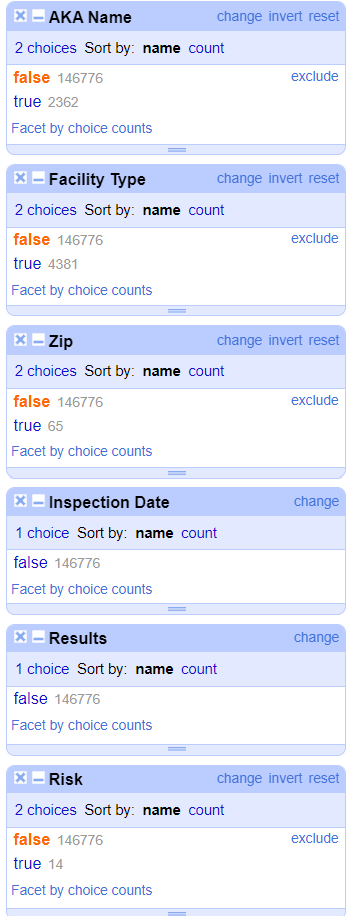


Table 3. Facets of cleaned columns

In this case, given the large size of the original dataset, which contains 150,000 observations, the reduction in the number of rows at this extent is deemed acceptable. Consequently, a preliminary dataset has been obtained that is well-suited for cleaning, as it no longer includes any empty values and has been restructured to facilitate further data cleaning.

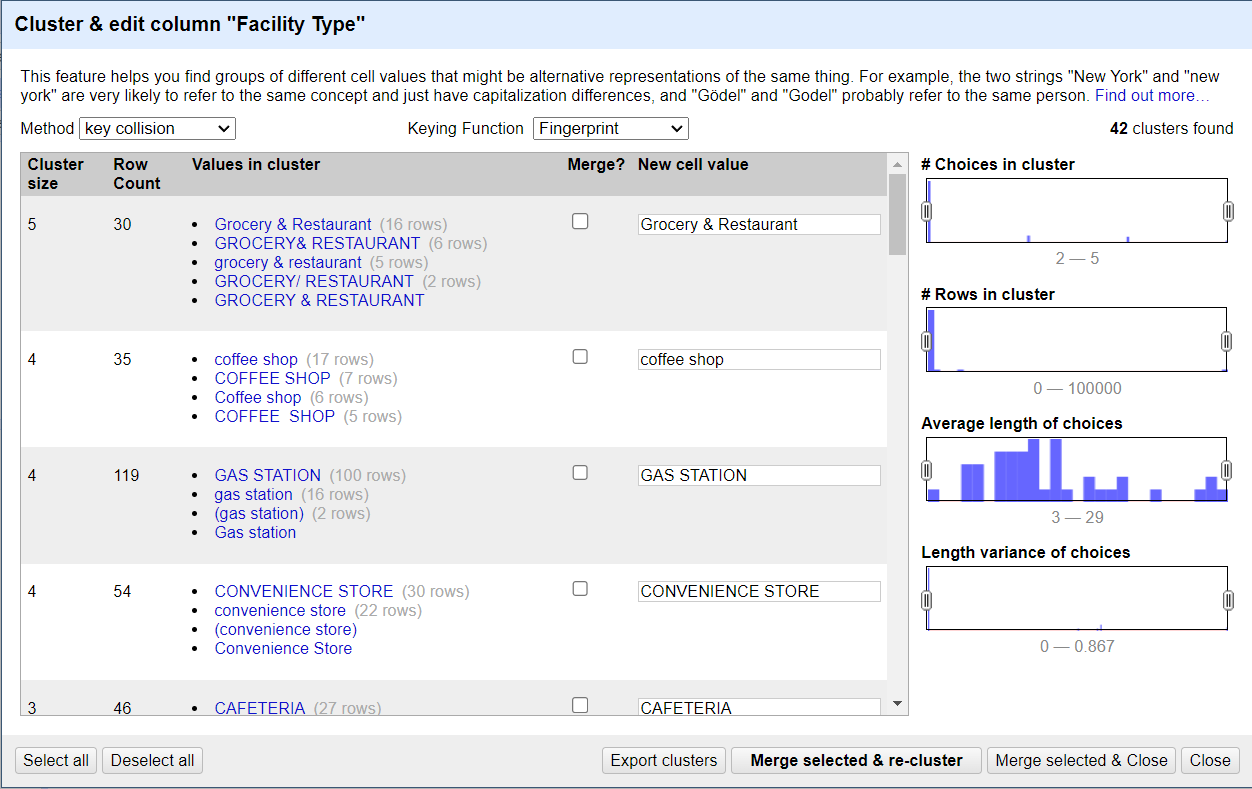
### 3.1.2 Step 2: Handle Facility Type

In use cases 1 and 2, the most frequently used column is Facility Type. Therefore, the results of data cleaning on Facility Type will have a significant impact on the subsequent data analysis.

In OpenRefine, I use [edit cells --> cluster and edit] to create groups by finding similar values in a column, making it easy to merge similar values into one group. This reduces data redundancy and improves data quality.

First, as shown in Figure#8, six different key functions are used to deal with similar and duplicates in the column of Facility Type [Fingerprint, ngram-fingerprint, metaphone3, cologne-phonetic, Daitch-Mokotoff, Beider-Morse].

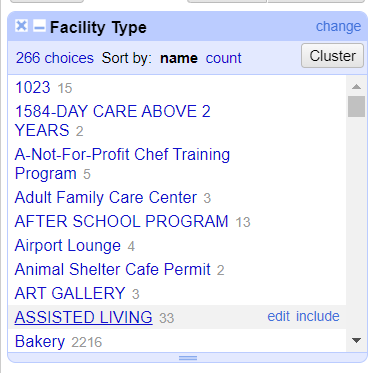
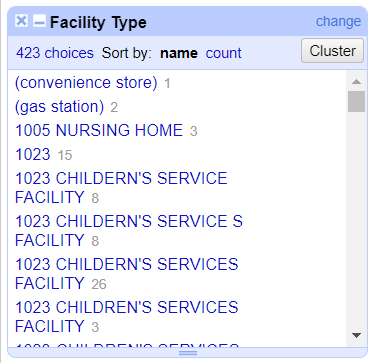
Figure 8. Clean Facility Type (cluster &edit)



When using the Cluster & Edit functionality, it is important to also use personal judgment to merge similar items. Be specific, when some automatic correction suggestions are posed, some manual changes need to be conducted. For example, [CONVIENCE STORE <--> CONVIENCE], [TAVERN <--> TAVERN-LIQUOR].

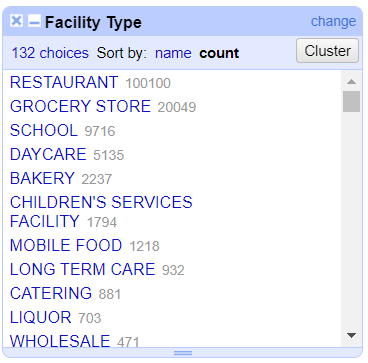
As shown in Figure#9, when this automatic cluster process is done, the Facility Type choices drop from 423 to 266.

Figure 9. Facet of Facility Type (automatic cleaning)



Then it is inevitable to clean the rest of 266 categories manually by comparing the similar meanings, functions and alphabet of each food supplier type. It is notified that before manually cleaning, it is better to transform the types into uppercase, so that it will reduce several spelling issues.

Figure 10. Facet of Facility Type (manually cleaning)



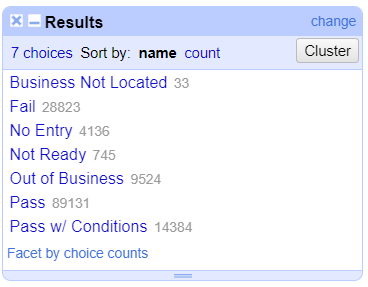
As shown in Figure#10, when the manually cleaning cluster process is done, the Facility Type choices drop from 266 to 132.

In this section, the initial cleaning of Facility Type is mostly completed. The result of the cleaning is that the elements in Facility Type have been reduced from 432 in the raw data to only 132.

### 3.1.3 Step 3: Handle Results

Based on Figure#11, it seems that there is no groups in Results that need to be clustered or merged, as all the categories are accurately divided into seven parts (Business Not Located, Fail, No Entry, Not Ready, Out of Business, Pass, Pass w/ Conditions).

Figure 11. Facet of Results



Based on the completed initial cleaning of Facility Type, the next step is to work on the objectives of use case 1. This section 3.3 focuses on how to calculate the passing rate of each facility type and visualize the data for comparison.

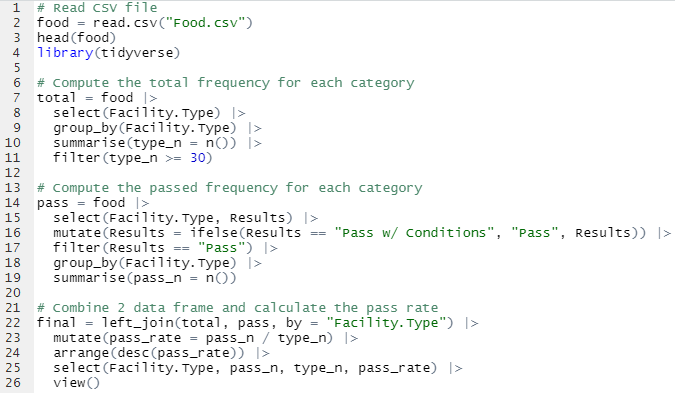
Next, export the cleaned dataset as a CSV format and save it. In R, read the CSV file and use the tidyverse package to analyze the target columns (Facility Type, and Results).

There are several crucial steps need to be notified:

1. The general approach is to obtain the frequency of passing inspections and the total frequency for each Facility Type from the data, and then calculate the “passing rate” by dividing the former by the latter.
2. Combine the "Pass" and "Pass w/ Conditions" statuses into a single "Pass" status for calculation purposes.
3. Imposing a minimum frequency threshold of 30 for each individual food supplier, which eliminates the majority of outliers and makes the comparisons statistically more meaningful.

The codes are listed in Figure#12, and the table output is shown in Table#4.

Figure 12. Data transformation on Facility Type and Results by using R (tidyverse)



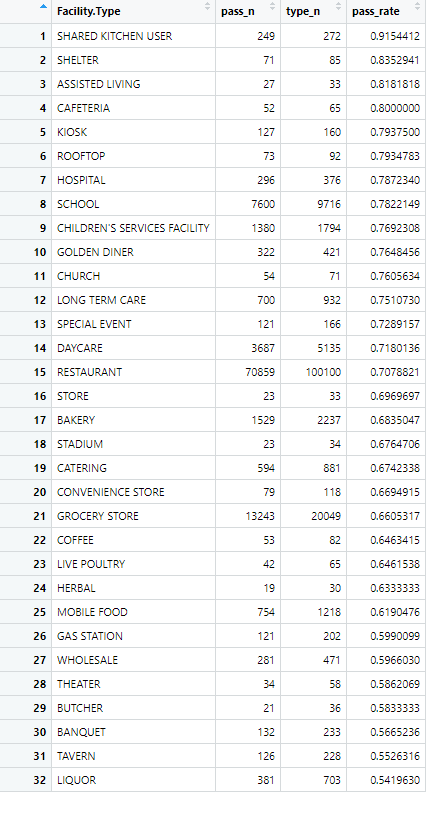


Table 4. Pass Rate of significant Facility Types

Based on the results, it can be observed that the final output of pass rates only covers 32 types of food suppliers, which is significantly fewer than the original total of 132. This is primarily due to the selection of food suppliers with a total frequency greater than 30, which resulted in the removal of most outliers and infrequent food suppliers. This filtering approach helps to focus the analysis on more common and meaningful cases.

Furthermore, it can be observed that Shared kitchen users have the highest pass rate, reaching up to 91.5%, while Liquor suppliers have the lowest pass rate at 54.2%.

### 3.1.4 Step 4: Handle Zip

Based on the previous analysis, it appears that the zip code does not require significant modifications. Therefore, similar to the approach used for Facility Type and Results, zip code can be grouped and summarized in R using similar methods. The resulting output is shown in Table#5.

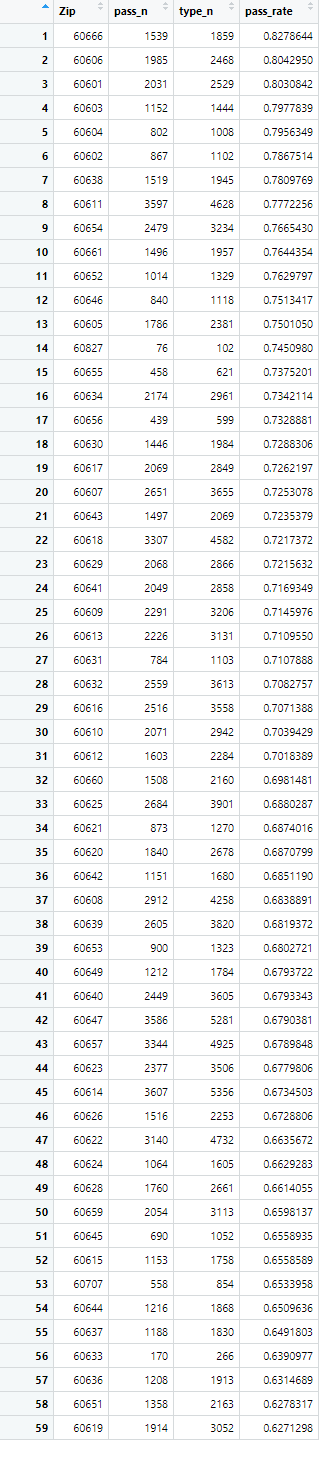
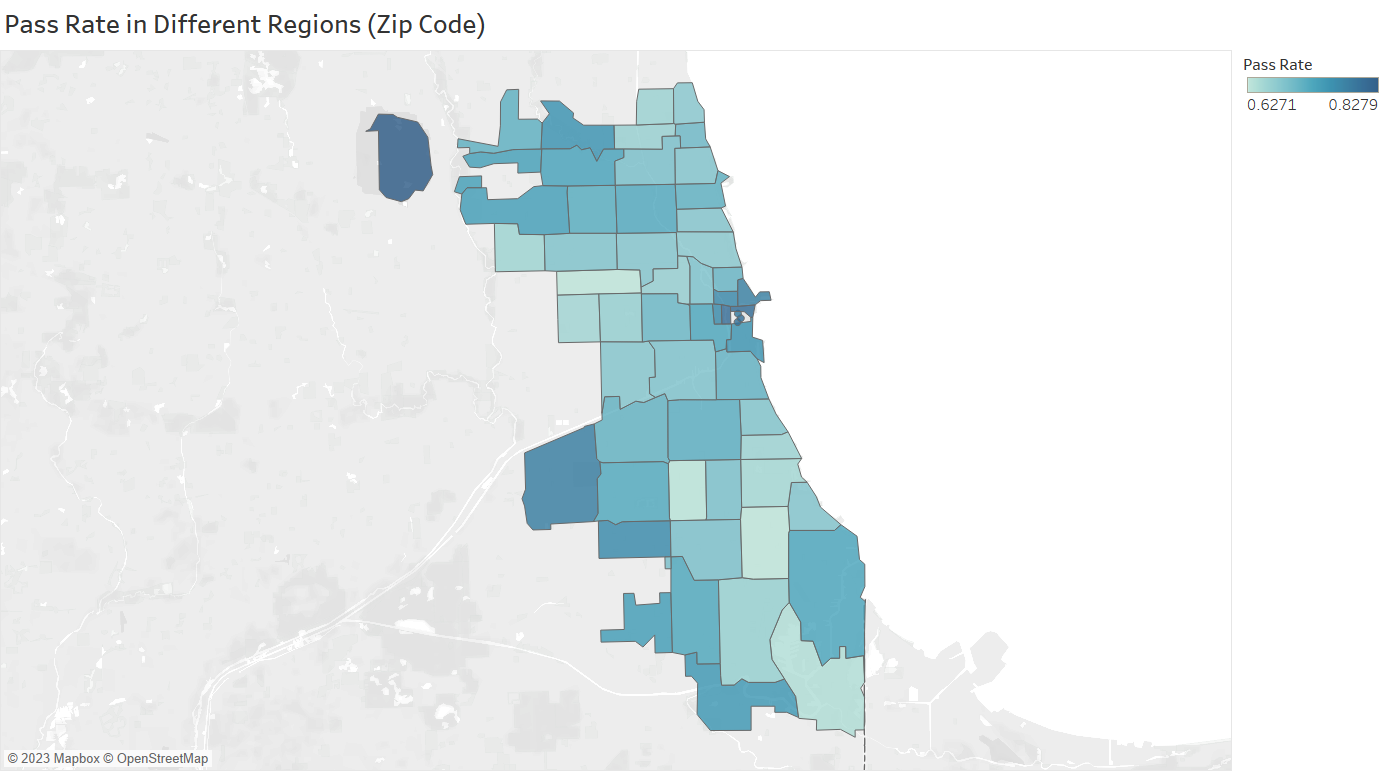


Table 5. Pass Rate in Different Regions (Zip)

Next, export the data from RStudio and use Tableau to identify zip codes and plot the geographical distribution of pass rate of food suppliers' food safety inspections on a map of Illinois based on their latitude and longitude coordinates (Figure#13).

Figure 13. Pass Rate in Different Regions (Zip Code)



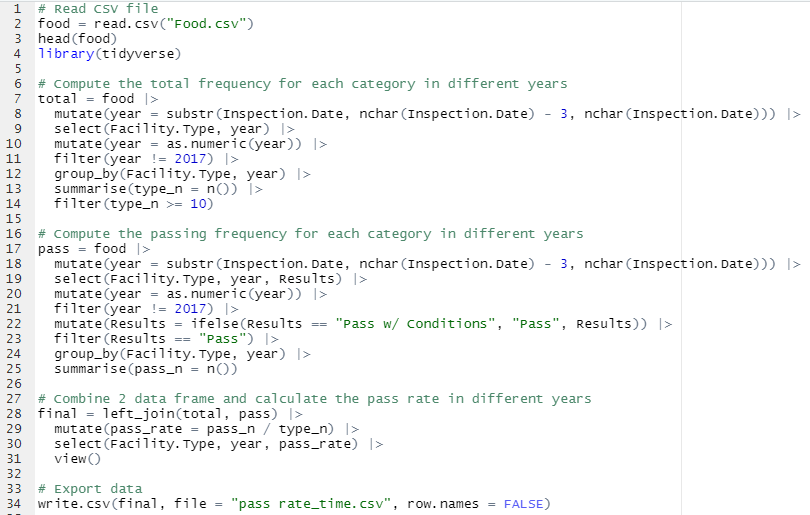
As shown in Table#4 and Figure#13, it appears that the region with zip code 60666 has the highest pass rate at 82.79%, while the region with zip code 60619 has the lowest pass rate at 62.71%. Additionally, the geographical distribution map can provide visual insights into the variation in pass rates across different regions. Darker colors indicate higher pass rates, while lighter colors indicate lower pass rates. It is observed that regions with higher pass rates tend to be geographically close to each other.

## Deal with Use Case 2

### Step 1: Handle Inspection Date

Based on the previous analysis, we can perform an analysis of the variation in pass rates across different facility types over time. The data transformation and processing can be done in R. The code for the operation is included in Figure#14, and there are several points that need to be explained:

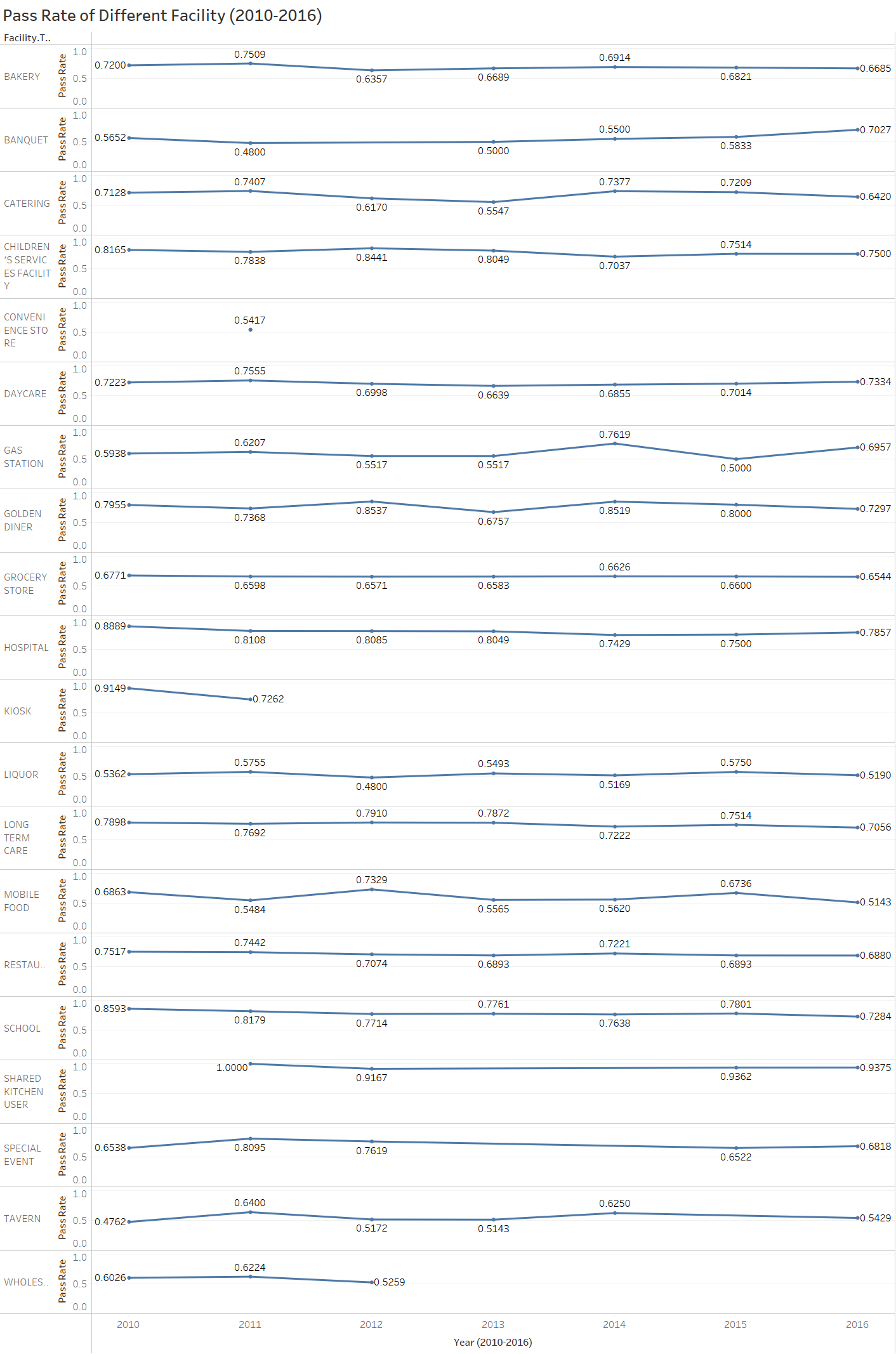
Figure 14. Data transformation on Facility Type, Inspection Date and Results by using R



1. In this Use case, the analysis mainly focuses on the yearly data. Therefore, it is necessary to convert the Inspection Date data type, which is character type, into integer type year data, which will facilitate the subsequent statistical analysis and visualization.
2. Since the time data for 2017 is not a complete annual data, it may have an impact on the results, so all data from 2017 is excluded.
3. To ensure the statistical significance of the data and eliminate possible outliers, the number of inspected food suppliers is controlled to be greater than or equal to 10 for each year.

After all these steps are completed, Tableau is used for data visualization to observe the changes in pass rates for each Facility Type over time. The line charts are shown in Figure#15.

Figure 15. Pass Rate of Different Facility (2010-2016)



The Figure#15 clearly show the trends of pass rates over time for each Facility Type, providing insights into which types have seen improvements or deteriorations over the years.

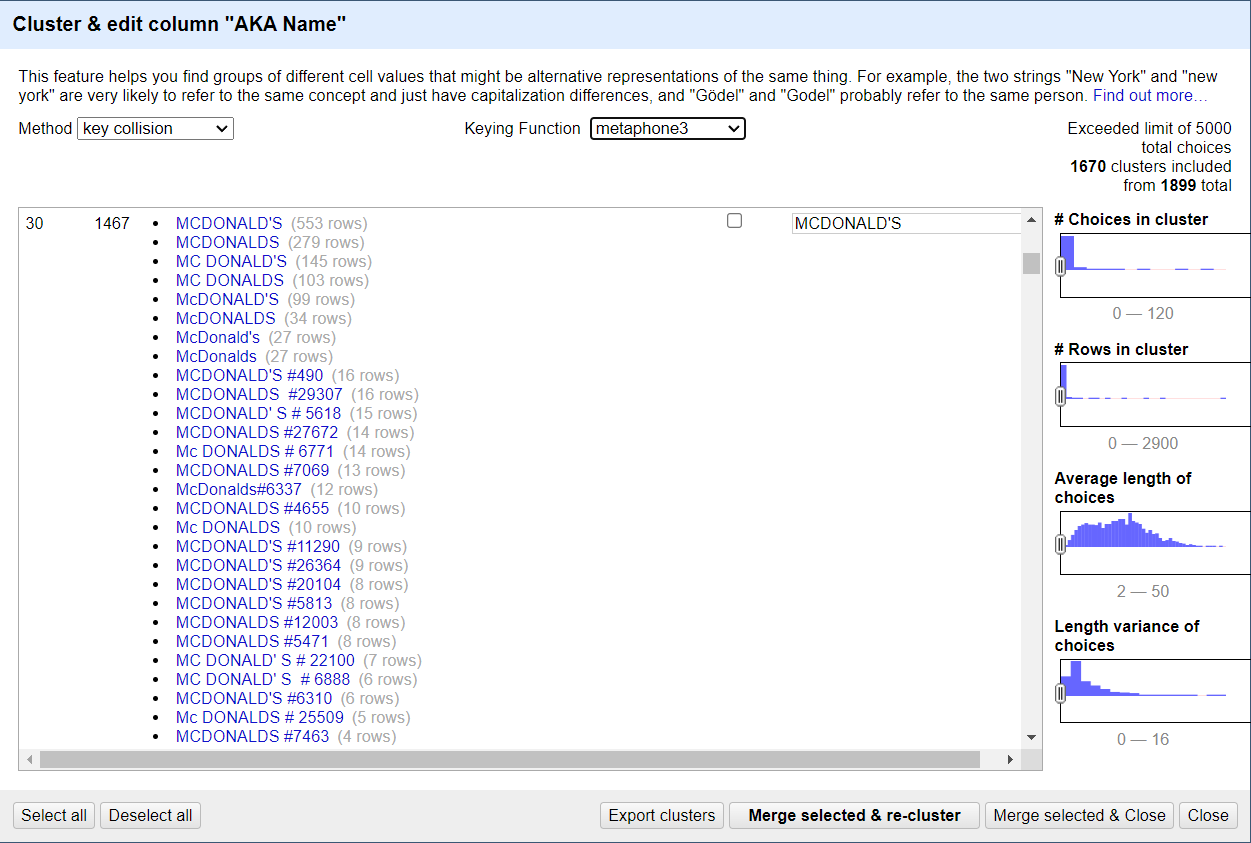
In addition, it can be observed that there is a small portion of types with missing annual data. There may be several reasons to explain this situation: it may be due to the fact that the number of surveyed facility types per year needs to be greater than 10, so these data are discarded. Another possibility is that these data themselves have no annual data and have been deleted due to null values during data cleaning.

### Step 2: Handle AKA Name

The next step involves analyzing the pass rate of a specific food supplier by cleaning and filtering its AKA Name. In Use Case 2, McDonald's is chosen for an in-depth research of its pass rate.

Firstly, the "AKA Name" column is cleaned using the "cluster & edit column" function in OpenRefine. It should be noted that many McDonald's in the "AKA Name" column have unique identifiers attached to them, which is likely due to inspections being conducted at different locations. However, they all belong to McDonald's, so the different identifiers are ignored and all are categorized as one type: McDonald's. This issue is resolved by using the Key Function: metaphone3 as shown in Figure#16.

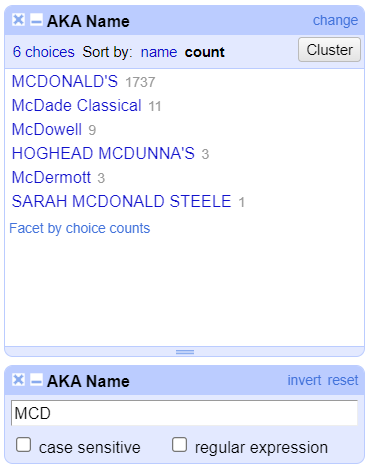
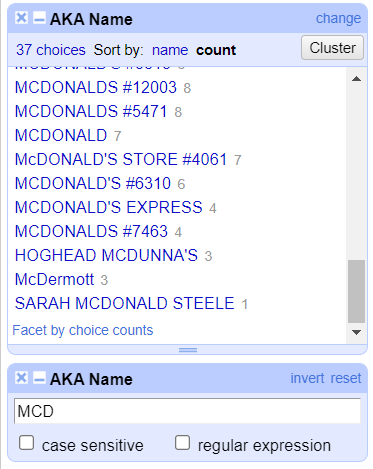
Figure 16. Cluster & edit column with “metaphone3” (McDonald’s)



In addition to the metaphone3 key function, there are five other key functions that can be used to further cluster the AKA Name column: Fingerprint, ngram-fingerprint, cologne-phonetic, Daitch-Mokotoff, and Beider-Morse.

After the automatic cleaning process, further refinement is made by using the text filter to merge McDonald's more precisely. As shown in Figure#17, the number of choices decrease from 37 to 6 after the final processing. The target food supplier, McDonald's, appeared 1737 times in the dataset.

Figure 17. Facet of AKA Name (McDonald’s)



When the new cleaned dataset is updated, then the final code used for data transformation and visualization in R is shown in Figure#18. The pass rate of all inspected McDonald's stores is shown in Figure#19.

Figure 18. Data transformation on AKA Name (McDonald’s) in R

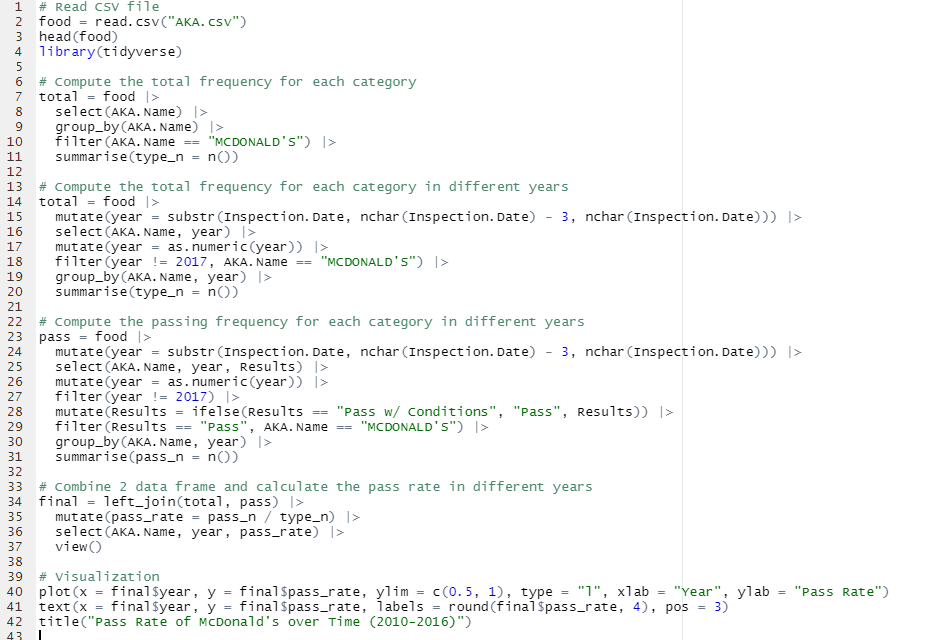
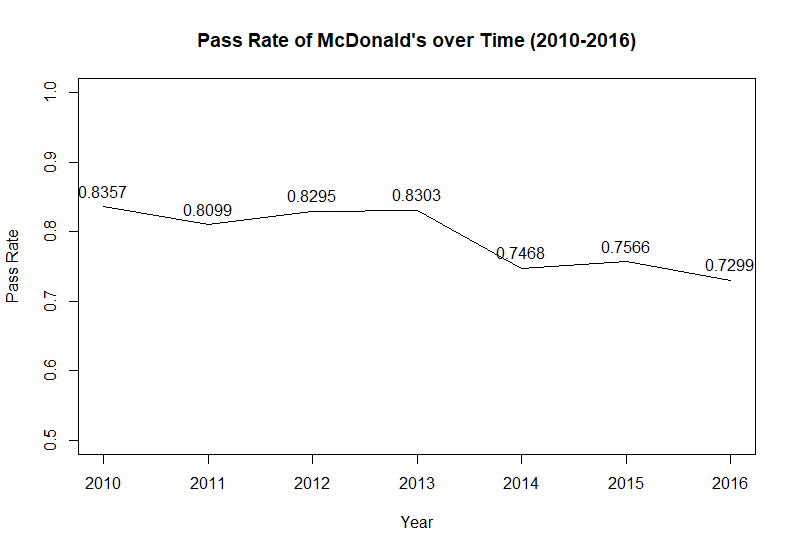
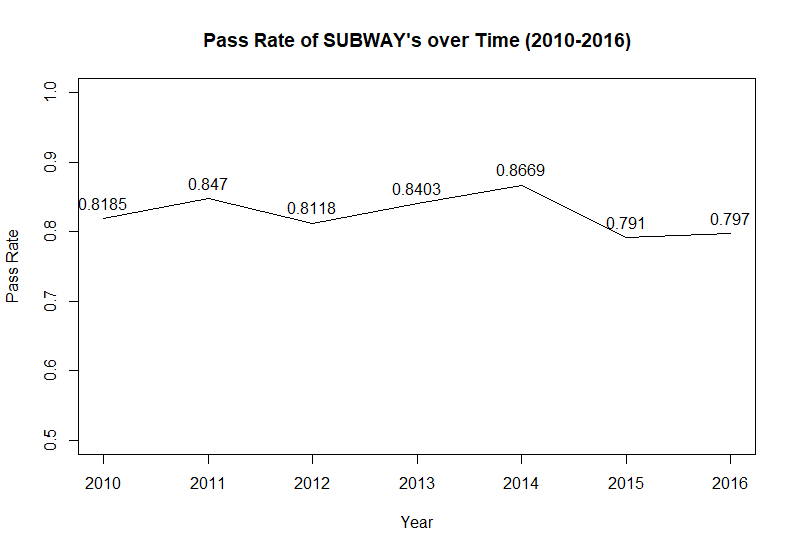


Figure 19. Data transformation on AKA Name (McDonald’s) in R



According to the Figure#18, it can be observed that the pass rate of inspected McDonald's gradually decrease from 2010 to 2016. Especially from 2013 to 2014, the pass rate experienced a dramatic drop.

Figure 20. Data transformation on AKA Name (SUBWAY) in R



The pass rate of Subway has been relatively stable over the past seven years. Since the facility type of Subway and McDonald's both belong to the restaurant category, it can be speculated that the reason for the decline in McDonald's pass rate from 2013 to 2014 may not be due to stricter inspection, but may be more due to internal problems of McDonald's itself.

# Evaluation

• Pass rate v.s. Results: By comparing Figure#2 and Figure#13, it can be observed that the range of pass rates has changed from 0.1667~1.0 to 0.6271~0.8279. Additionally, on the map, several isolated areas have disappeared, and most of the surveyed areas are concentrated in a region in the western part of Illinois. This indicates that areas containing extreme values have been excluded. From Figure#1, it can be roughly inferred that observations with fewer surveys in the Zip codes between 60007 and 60559 are discarded, and only statistically significant areas with high pass rates are displayed.

• Facility Type v.s. Results: By comparing Table 1 and Table 4, it can be observed that the cleaned pass rate ranges from 54.20% to 91.54%, and a large amount of data with a pass rate of 1.0 has been removed from the original data. Additionally, all null values have been deleted, and there are no longer any instances of pass rate being displayed as NA.

• Results of Facility Type vs. Inspection Date: By comparing Figure#5 and Figure#15, it can be observed that classifying the data based on Results does not showcase the most significant changes in pass rate over time for facility types. However, Figure#15 is able to highlight the most significant changes in pass rate over time for the few selected facility types, better showcasing the underlying patterns in the dataset.

• AKA Name and Pass Rate: Comparing Figure#7, it can be observed that the selection of DBA Name or AKA Name can impact the subsequent data processing and the level of difficulty involved. However, by using AKA Name for analysis, as shown in Figure#17,#19, and #20, it can reduce the difficulty of further classification to a certain extent. Additionally, it can improve the accuracy of targeted time-based analysis for certain food suppliers.

# Refined Data Cleaning

The objective of the Refined Data Cleaning is to address specific issues identified during the Initial Data Cleaning. These include providing additional details regarding the clustering of Facility Types, such as the criteria for similarity determination and the specific clustering algorithm used. Additionally, a more comprehensive and transparent explanation of how to handled null values will be added, including the method used to handle missing data and the rationale behind it. Furthermore, further detailed explanations to other data cleaning process will be provided.

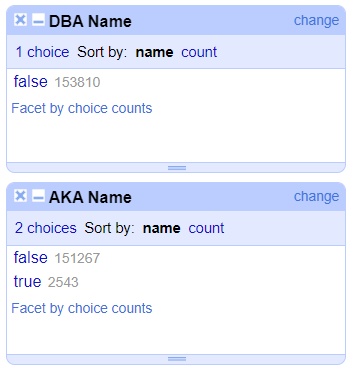
## Basic Cleaning

Unlike the Initial Data Cleaning, where only six relevant columns are kept based on Use Case 1 and Use Case 2, all columns of the original dataset are retained in the Basic Cleaning phase. This ensures that the data cleaning results are maximally reusable. To summarize, Facility Type, Results, Zip, AKA Name, Inspection Date, and Risk will still undergo cleaning, while other irrelevant columns will still be kept, which maintains the data's completeness and to allow for any additional analyses that may be relevant beyond Use Case 1 and Use Case 2.

In contrast to the Initial Data Cleaning, null values will not be handled during the Basic Cleaning phase. The process of handling null values will vary depending on the specific variables, which will be discussed in detail in the following sections. This decision is made to allow for a more targeted and flexible approach to dealing with missing data based on the specific characteristics of each variable, which can lead to more accurate and effective data cleaning results.

In addition, it should be noted that the handling of DBA Name and AKA Name differs in subsequent data processing. Specifically, only DBA Name will be utilized in the processing of use cases. An examination of the data using Facet by blank (null or empty string) (Figure#21) reveals that DBA Name has no null values, while AKA Name has 2,543 null values. Therefore, utilizing DBA Name to process use cases represents a superior approach, as it maximizes the completeness of the information.

Figure 21. Facet by blank on DBA and AKA Name (null or empty string)



Given that converting Inspection Date to a year format in R can be cumbersome, OpenRefine will be used to preprocess Inspection Date in this section in advance which can simplify the data cleaning process and reduce the workload required in R, which can help to streamline the workflow and improve efficiency. After observing all Inspection Dates in the dataset follow the "mm/dd/yyyy" format through Facet, the "split into several columns..." option will be used in the Edit columns tool in OpenRefine. This will split Inspection Date into three separate columns for "month", "day", and "year" using the "/" delimiter. The columns representing "month" and "day" will then be deleted, as they are not needed for the data cleaning process.

## 5.2 Facility Type Refined Cleaning

According to the Initial Data Cleaning, Facility Type is one of the most important variables in the Use Cases. The quality of its classification can determine the performance of Pass Rate in subsequent analyses. Therefore, in this section, specific cleaning approaches for Facility Type will be analyzed to ensure its accuracy and consistency.

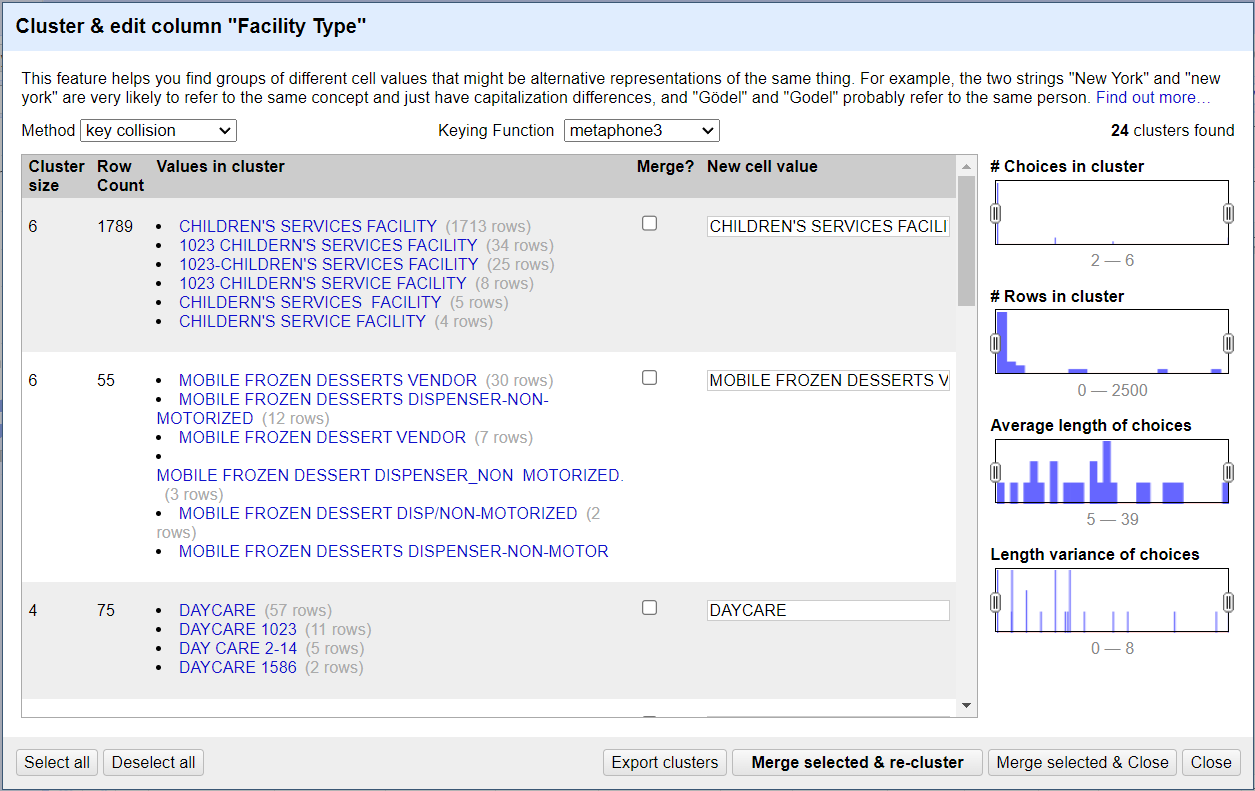
**Step 1:** **Capitalize Facility Type**

All the observations in Facility Type are converted to uppercase, which make it easier to cluster similar facility types.

**Step 2: Automatically Cluster & Edit Facility Type**

The 6 key functions in “Cluster & Edit” are used to automatically cluster Facility Type. The cleaning procedures are conducted by the order of the functions: Fingerprint, ngram-fingerprint, metaphone3, cologne-phonetic, Daitch-Mokotoff, Beider-Morse.

Figure 22. Cluster & Edit on Facility Type (metaphone3)



Metaphone3 is a phonetic algorithm that can help to identify and group similar sounding words together, which can be useful for clustering Facility Types with common pronunciation.

For instance, as shown in Figure#22, when clustering Facility Type for "DAYCARE", other observations with similar pronunciations containing the word "DAYCARE" will be included in the comparison. In the possible types of Facility Types, many similar elements are simply numbers added after "DAYCARE". These numbers may represent the age group that the daycare serves (2-14 years old), or possible store IDs (1023 & 1586). In fact, these variations are all belong to “DAYCARE”, which can be merged together as one single category “DAYCARE”. This principle will be applied to the whole data cleaning process, which means the additional information of a specific Facility Type should be omitted and all these types should be merged into one specific Facility Type. For example, “DAYCARE (2-6 YEARS)” should be clustered as “DAYCARE”.

Figure 23. Assumption of Cluster & Edit on Facility Type (metaphone3)

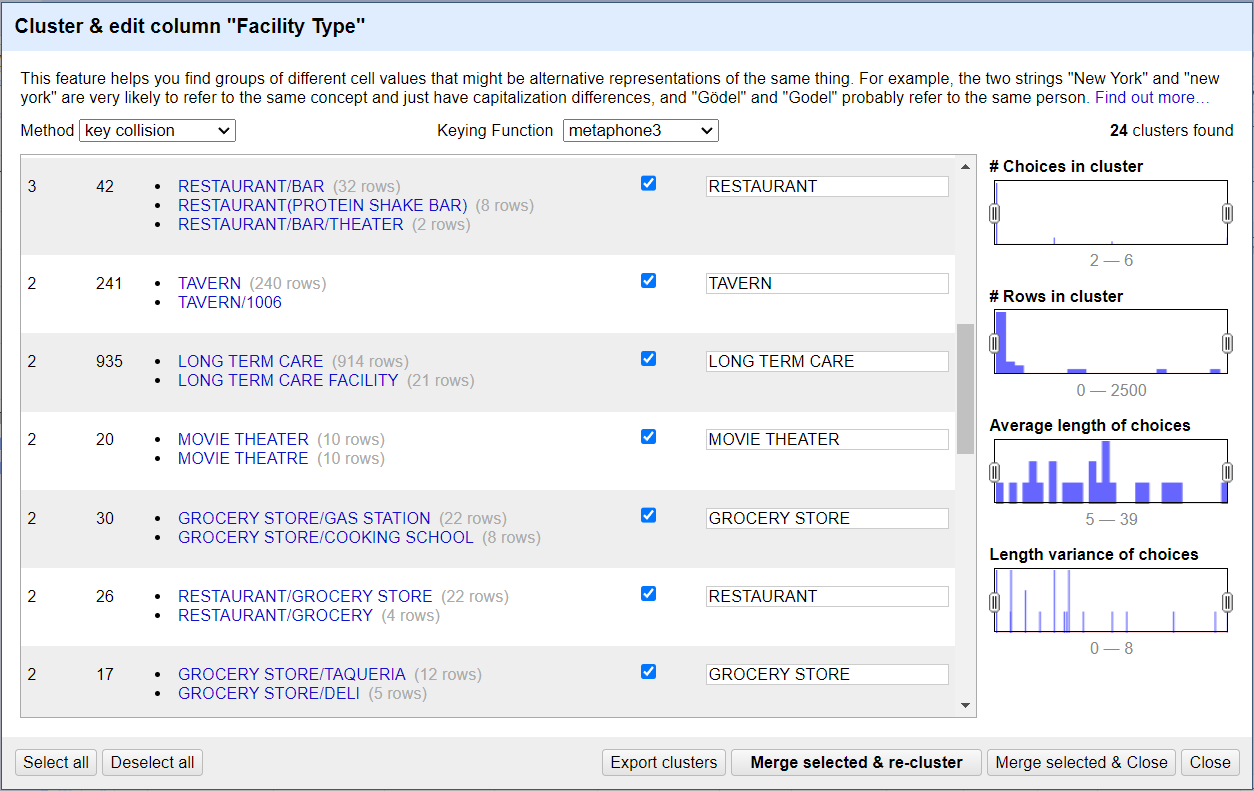
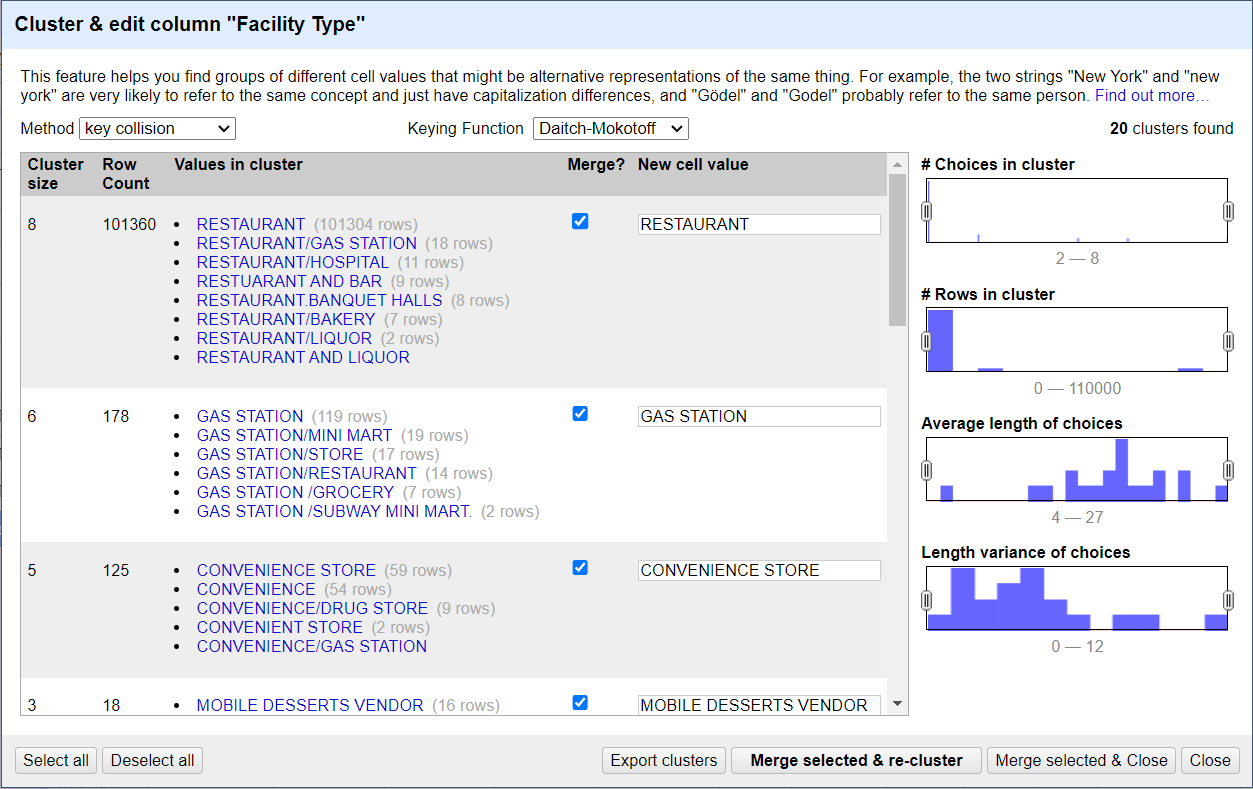


Figure 24. Assumption of Cluster & Edit on Facility Type (Daitch-Mokotoff)



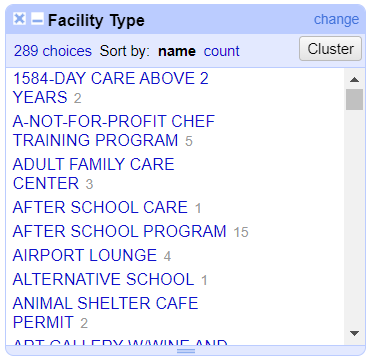
As shown in Figure#24, it may be observed that an observation of Facility Type is a new type combination of two or more specific food supplier types separated by a "/" or combined by “AND” (e.g. RESTAURANT/BAR/THEATER). In such cases, it is hypothesized that the cluster strategy should be based on the first type listed in the combination type. For instance, "RESTAURANT/BAR/THEATER" should be clustered as "RESTAURANT".

After the completion of Automatic Cluster & Edit, the Facility Type is reduced from the original 447 to 289.

**Step 3: Deeper Cluster & Edit Facility Type**

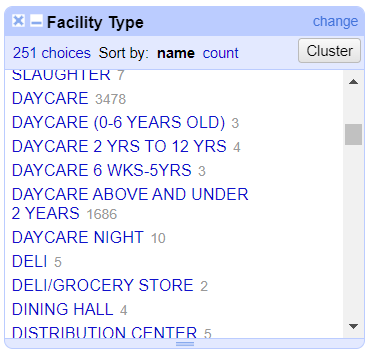
The similar Facility Types can be merged manually by observing the text facet sorted alphabetically.

Figure 25. Facet of Facility Type for manual



For example, in Figure#25, "1584-DAY CARE ABOVE 2 YEARS 2" can be grouped into "DAYCARE". "AFTER SCHOOL CARE" and "AFTER SCHOOL PROGRAM" are very similar and can both be grouped into "AFTER SCHOOL PROGRAM", which has a higher frequency (15 > 1).

Figure 26. Facet of Facility Type for manual



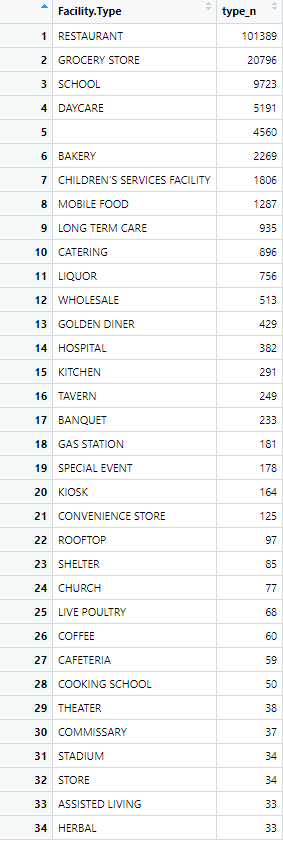
Furthermore, it can be observed from the sorted text facet that some Facility Types have many similar types, which only differ by a part of descriptive words. However, these types actually belong to the same category. Therefore, they can be merged into one category. For instance, in Figure#26, all names of types beginning with "DAYCARE" can be categorized under "DAYCARE".

After completing the Deeper Cluster & Edit Facility Type operation, the final result of manually merging similar types reduces the number of Facility Type choices from 289 to 140.

**Step 4: Cluster Facility Type with Null values**

To continue the data cleaning process, the next step is to address Facility Types with null values. After completing Steps 1-3, it is observed that there are still 4560 observations with blank values. Therefore, it is necessary to handle these nulls while minimizing the risk of introducing classification errors.

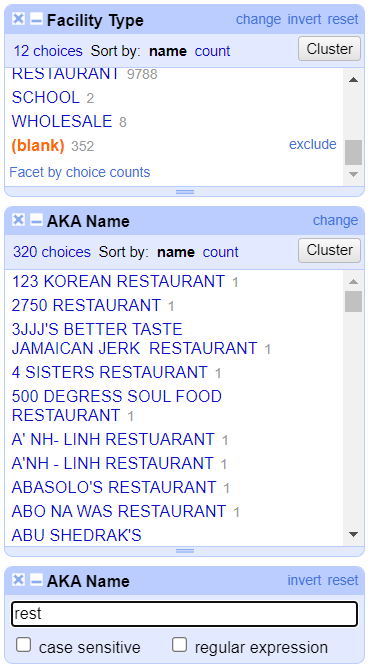
Figure 27. Overview of the frequencies of Facility Type (type\_n ≥ 30)



Firstly, select Facility Types with a frequency of 30 or higher in R and sort them in descending order by frequency, including null values. Then, 34 types are filtered. Then, null values can be filled by determining Facility Types based on these higher frequency types, while minimizing classification errors.

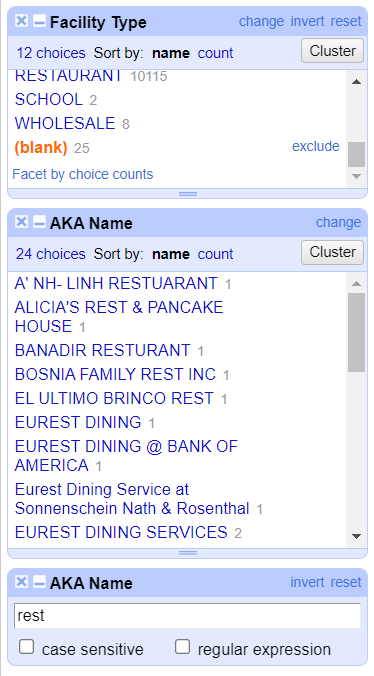
Next, the information provided by DBA Name can be used to fill in the null values. The main approach is to use text filters to search for keywords in DBA Names in sequence, as shown in Figure#28, in order to quickly fill in the blank values.

Figure 28. Deal with null values by Text filter



For example, starting with the most frequent Facility Type, which is RESTAURANT, all blank values under this type are included. Then, the Text filter is used to input the keyword "restaurant". This yields 352 null values that fit this description. By combining this information with the DBA Name column, it becomes clear that there are many different types of restaurants. Therefore, a manual check is conducted to ensure that the majority of Facility Types within this group can be classified as RESTAURANT.

Figure 29. Deal with null values by Text filter (key word of DBA Name)



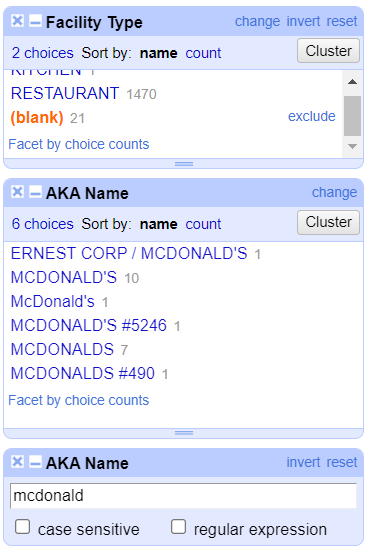
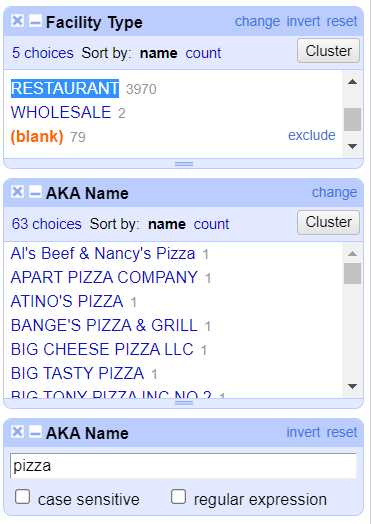
After the cleaning of "restaurant", the keyword can be shortened to "rest". As shown in Figure#29, there are some food suppliers with the abbreviation "REST" (e.g. ALICIA'S REST & PANCAKE HOUSE), as well as some spelling errors (e.g. BANADIR RESTURANT). Therefore, the Facility Types corresponding to these DBA Names can also be determined.

When finish cleaning “RESTAURANT”, the null values drop from 4560 to 4211.

When all the cleaning procedures done, the total number of null values drop from 4560 to 3393.

The deep cleaning involves either inputting certain keywords related to the services provided by the food suppliers or checking the names of some well-known food suppliers. In Figure#30, it appears that the text filter based on DBA Name can recognize observations that contain the keywords "pizza" or "mcdonald" in the food supplier's name, and classify null values related to these keywords as "RESTAURANT".

Figure 30. Deal with null values by Text filter (key word of service & well-known supplier)



After the manual cleaning process, the null values in Facility Type drop from 4560 to 1206. The clustered Facility Type contains 133 choices.

## 5.2 Other Refined Cleaning

The DBA Name can be cleaned in a similar way alike what has been done on AKA Name in 5.2 Facility Type Refined Cleaning. The main target food suppliers are McDonald’s and Subway, which is researched in Section 3.2.2.

Since Inspection Date has been processed in Section 5.1, thus there is no need to clean this factor any more.

According to the facets, there are no nulls in Results. The nulls in Risk (85 observations) and Zip (98 observations) are removed from the dataset, which make up an extremely small percentage of the data of up to 150,000. Thus, they are simply removed from the data in order to reduce clustering errors and ensure that the dataset performs well.

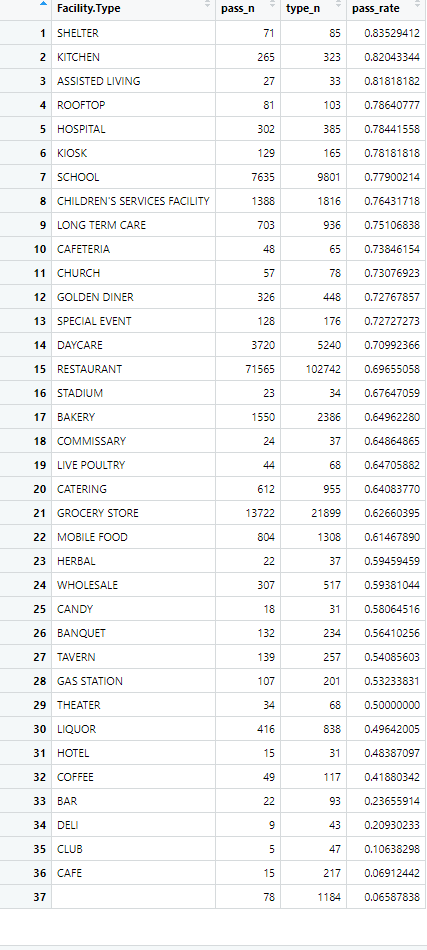
Finally, Refined Data Cleaning is done and the total observations drop from 153810 to 153627.

# Data Visualization and Evaluation

## Pass Rate of Facility Type

In the Refined Data Cleaning process, the significantly high Pass Rate of Facility Type count increased from 32 to 37 (including the unprocessed null values) after the final calculations are performed. It can be observed that, apart from some newly added facility types, the ranking of the overall pass rate remain largely unchanged.

Figure 31. Pass Rate of Facility Type



The Pass Rate of Facility Type over time is shown in Figure#32. After the Refined Data Cleaning process, there are 35 significantly changing pass rates of facility type over time, which is a significant improvement over the 20 observed in the Initial Data Cleaning. However, there still exist certain facility types that only exhibit a significant pass rate in certain years or for a few years. After filtering out some of these locally significant facility types (Figure#33), the count of significantly changing pass rates of facility type over time decrease from 35 to 25, which is still an improvement over the results obtained in the Initial Data Cleaning.

Figure 32. Pass Rate of Facility Type over time (unfiltered)

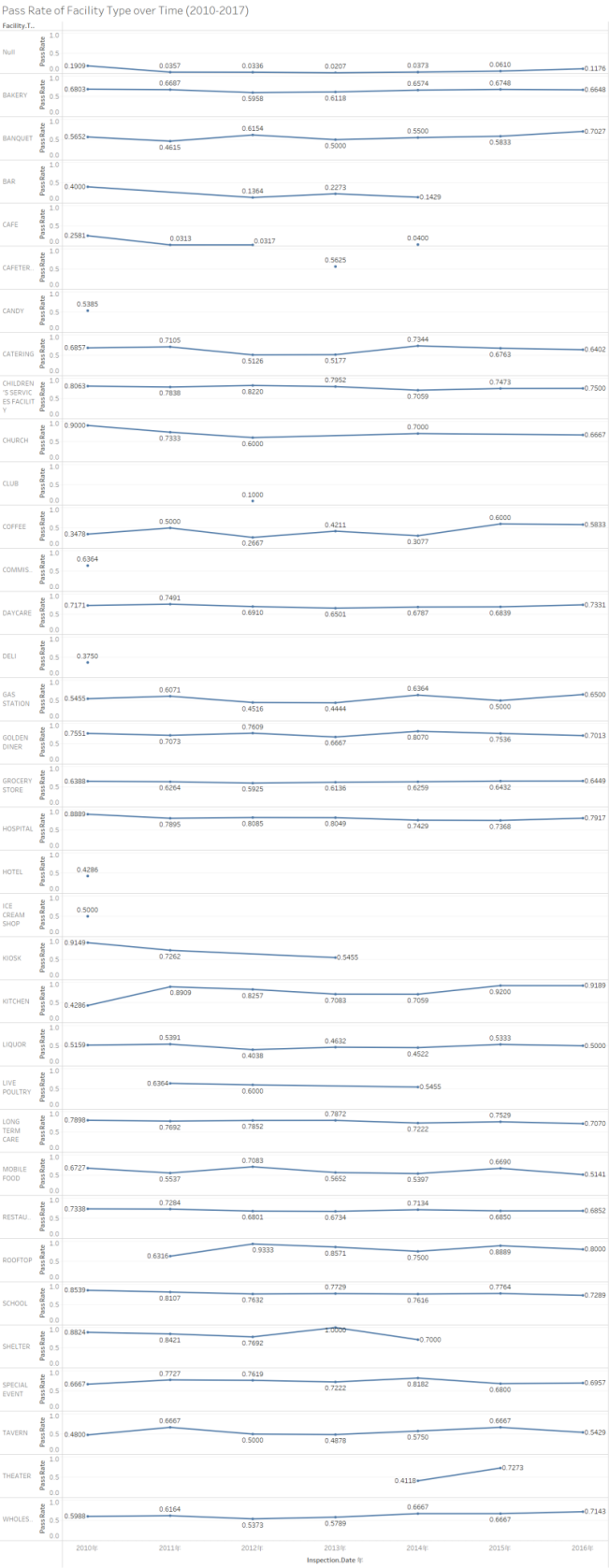
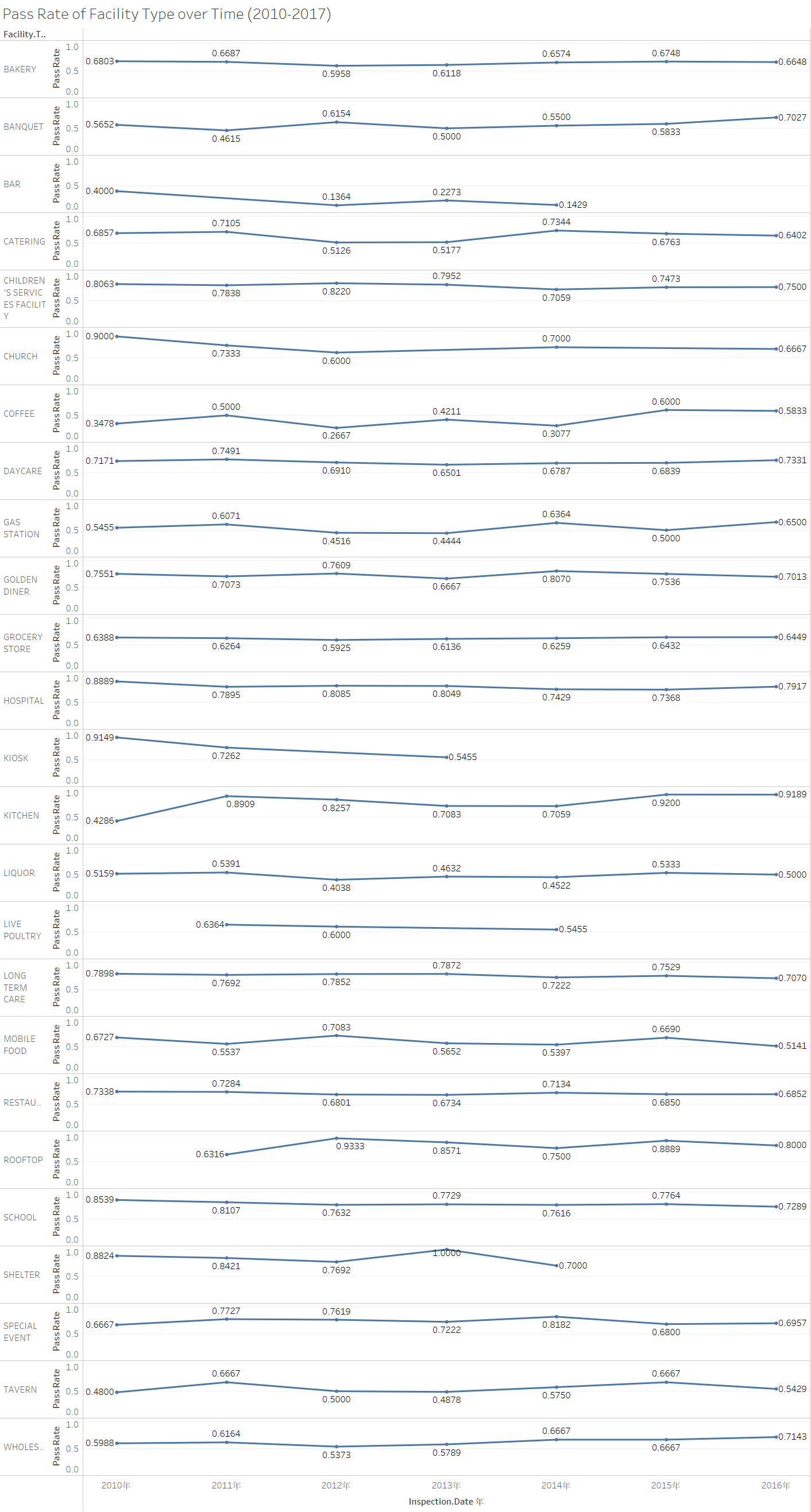


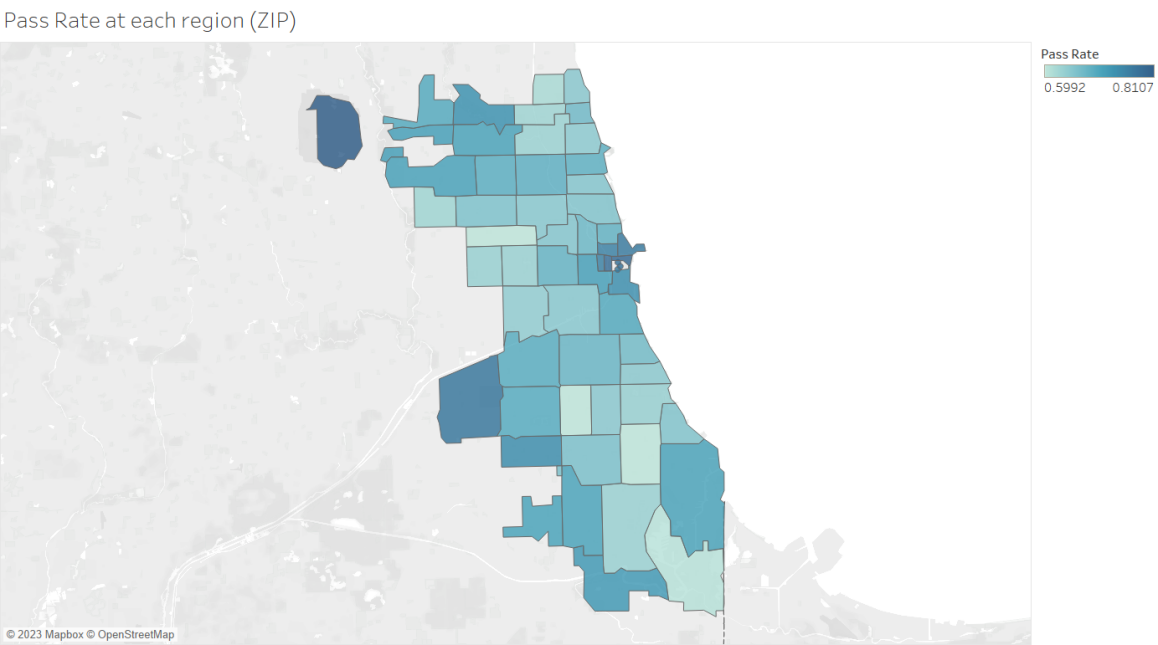
Figure 33. Pass Rate of Facility Type over time (filtered)



## Pass Rate at Each Region (Zip code)

Figure#34 displays the Pass Rate at Each Region based on the refined cleaned Zip data. Compared to the results obtained from the initial cleaning, there are no significant changes observed.

Figure 34. Pass Rate at different regions



## Pass Rate of McDonald’s & Subway over Time

Figure#35 and #36 display the Pass Rates of McDonald’s and Subway over time. Compared to the results obtained from the initial cleaning, there are no significant changes observed.

Figure 35. Pass Rate of McDonald’s over time (2010-2016)

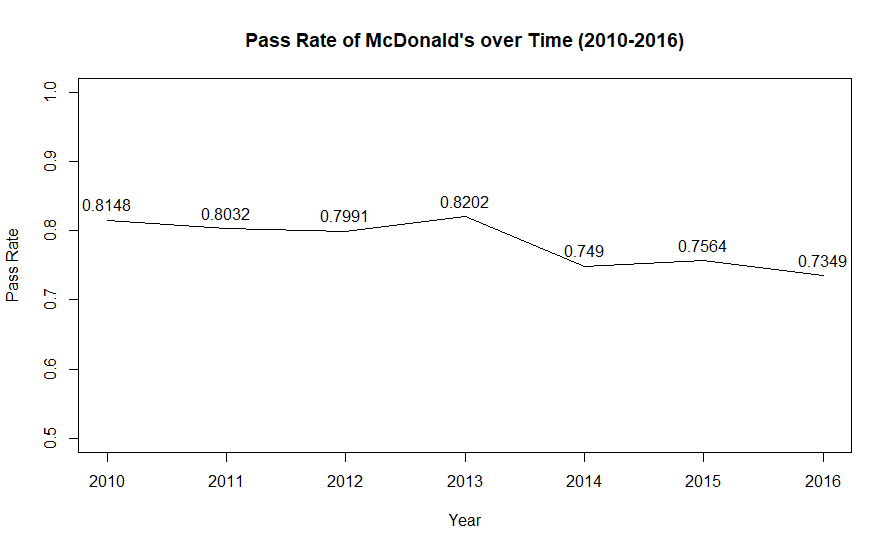
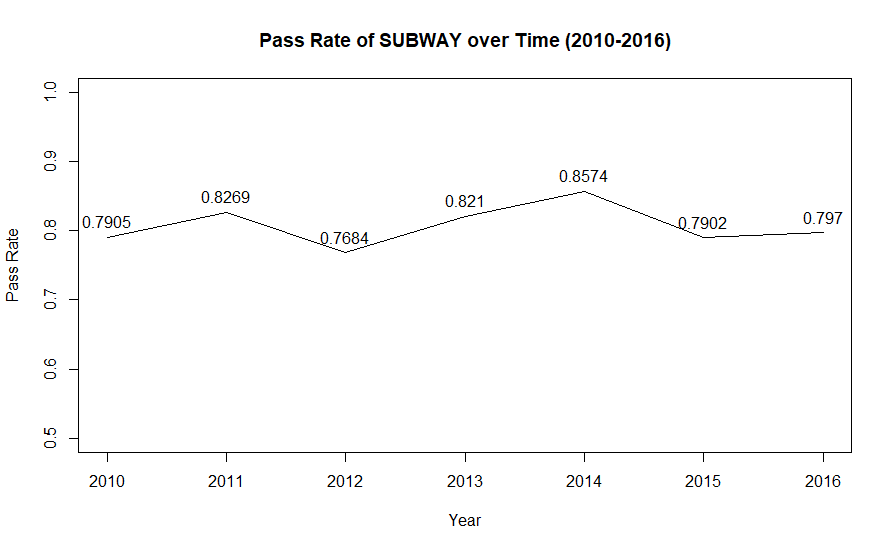


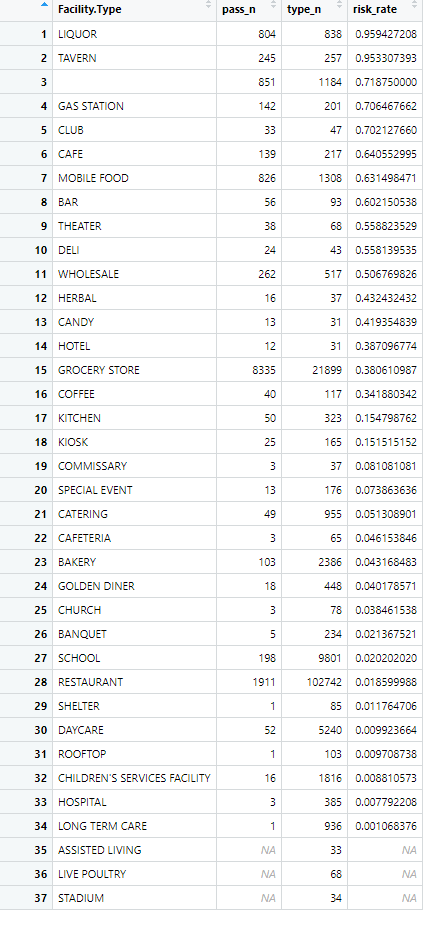
Figure 36. Pass Rate of Subway over time (2010-2016)



## Risk Rate of Facility Type

The calculation presented here represents the percentage of each Facility Type that has been rated as Risk 3: Low. It can be observed that there are 34 Facility Types with a significantly high risk rate. The Facility Type with the highest Risk Rate (Low) is LIQUOR, while the Facility Type with the lowest Risk Rate (Low) is LONG TERM CARE.

Figure 37. Risk rate of facility type (risk: Low)



Based on these results in Figure#37, it can be inferred that the determination of Risk is not solely based on the Pass Rate of restaurants but rather involves a comprehensive evaluation of the types of facilities and their involvement in food safety practices. This is why the output of these results is so different from the Pass Rate of Facility Type presented in section 6.1.

# Conclusion

This article primarily describes the Food Inspection Records conducted by CDPH from 2010-2017. The article focuses on two use cases, where the columns AKA Name, DBA Name, Facility Type, Risk, Zip, Inspection Date, and Results undergo initial and refined data cleaning, followed by data analysis and visualization. Based on the results of the refined data cleaning, the total valid observations decrease from 153,810 to 153,627. The most significant factor, Facility Type, decrease from 423 to 133. Regarding food/sanitation safety pass rates, it is observed that SHELTER had the highest pass rate, while LIQUOR had the highest probability of being rated as Risk 3 (Low). 25 Facility Types show significant pass rate changes over the years, and 34 Facility Types show significant Risk 3 (Low) rate. Some counties in Illinois are also visualized based on their corresponding pass rates, with darker colors indicating higher pass rates. Two restaurants, McDonald's and Subway, are selected and cleaned based on their DBA Name to observe the trend of their pass rates over time. The pass rate of McDonald's shows a slightly decreasing trend, while Subway's pass rate remains relatively stable.

The fully cleaned CSV file was saved and named "Food-Inspections-csv 4.27", which contains a total of 15627 observations and 17 columns. Only the AKA Name, DBA Name, Facility Type, Risk, Zip, Inspection Date, and Results columns were cleaned. Based on this file, analysis related to the six target columns use cases can be conducted or further cleaning can be conducted.

# Appendix

## Summary Table

### 1.1 Initial cleaning summary

|  |  |  |
| --- | --- | --- |
| **Initial Cleaning** | **Process** | **Details** |
| **Deal with Use Case 1** | | |
| Step 1: basic cleaning | 1. Remove irrelevant columns and reserve the targeted columns. | • OpenRefine  • 11 columns changes  • Human justification |
| 1. Reorder the left columns in order to make dataset understandable and easier to clean. | • OpenRefine  • 1 column changes  • Human justification |
| 1. Delete all null values. | • OpenRefine  • 7034 changes  • Automatically |
| Step 2: Handle Facility Type | 1. Cluster and edit similar and duplicate components | • OpenRefine  • Automatically  • Human justification |
| 1. Transform the name of Facility Type into uppercase | • OpenRefine  • 138261 changes  • Automatically |
| 1. Merge similar components manually | • OpenRefine  • 4065 changes  • Human justification |
| Step 3: Handle Results | 1. Combine the "Pass" and "Pass w/ Conditions" statuses into a single "Pass" status | • R  • 14384 changes  • Automatically |
| 1. Imposing a minimum frequency threshold of 30 for each individual food supplier | • R  • 100 changes  • Automatically |
| Step 4: Handle Zip | 1. Combine the "Pass" and "Pass w/ Conditions" statuses into a single "Pass" status | • R  • 14384 changes  • Automatically |
| 1. Imposing a minimum frequency threshold of 30 for each Facility Type from each Zip area | • R  • 98 - 59 = 39 Zip changes  • Automatically |
| 1. Visualization based on Zip and Pass Rate | • Tableau  • Automatically |
| **Deal with Use Case 2** | | |
| Step 1: Handle Inspection Date | 1. Combine the "Pass" and "Pass w/ Conditions" statuses into a single "Pass" status | • R  • 14384 changes  • Automatically |
| 1. Change Inspection specific Date to year only | • R  • Automatically |
| 1. Drop the 2017 data | • R  • Automatically |
| 1. Imposing a minimum frequency threshold of 10 for each Facility Type in each year | • R  • Automatically |
| 1. Visualization based on Year and Pass Rate | • Tableau  • Automatically |
| Step 2: Handle AKA Name | 1. Cluster and edit similar and duplicate components | • OpenRefine  • 20075 + 6865 = 26940 changes  • Automatically |
| 1. Merge similar “McDonald’s” manually | • OpenRefine  • 986 changes  • Human justification |
| 1. Data Transformation for data analysis | • R  • Automatically |
| 1. Data Visualization | • R  • Automatically |

### 1.2 Defined cleaning summary

|  |  |  |
| --- | --- | --- |
| **Refined Cleaning** | **Process** | **Details** |
| **Basic Cleaning** | | |
| Basic Cleaning | Transform the format of Inspection Date from “mm/dd/yyyy” to “year” only | • OpenRefine  • 153810 changes  • Automatically |
| **Use Case 1 & 2** | | |
| Step 1: Facility Type Refined Cleaning | 1. Capitalize Facility Type | • OpenRefine  • 146729 changes  • Automatically |
| 1. Cluster & Edit Facility Type | • OpenRefine  • Fingerprint: 1967; ngram-fingerprint: 2100; metaphone3: 6872; cologne-phonetic: 9; Daitch-Mokotoff: 125173; Beider-Morse: 0  • Automatically |
| 1. Merge similar components manually | • OpenRefine  • 4362 changes  • Human justification |
| 1. Based on the frequency order, using Text filter to fill out the blanks/nulls | • OpenRefine  • 3524 changes  • Human justification |
| Step 2: AKA Name Refined Cleaning | 1. Capitalize Facility Type | • OpenRefine  • 9585 changes  • Automatically |
| 1. Cluster & Edit Facility Type | • OpenRefine  • Fingerprint: 13436; ngram-fingerprint: 7328; metaphone3: 31365;  Beider-Morse: 1017  • Automatically |
| 1. Merge similar components manually | • OpenRefine  • 2275 changes  • Human justification |
| Step 3: Other Refined Cleaning | 1. Drop the misunderstanding values and nulls in Risk | • OpenRefine  • 85 changes  • Automatically |
| 1. Drop the nulls in Zip | • OpenRefine  • 98 changes  • Automatically |

## Overall Workflow

• **Tool:** MindMaster

