

# *INFO 399: Data Cleaning Research Project*

*Kaleigh Spitzer*

## Table of Contents

<b>1. Project Description .....</b>	<b>2</b>
1.1 Dataset Description .....	2
1.2 Use Case.....	2
1.3 Target Columns .....	2
<b>2. Process Overview .....</b>	<b>2</b>
<b>3. Execute on the Dirty Dataset .....</b>	<b>3</b>
<b>4. Data Errors and Quality Problems .....</b>	<b>5</b>
4.1 Typos.....	5
4.2 Non-Uniform Casing .....	5
4.3 Blank Entries.....	5
4.4 Formatting.....	5
<b>5. Initial Cleaning.....</b>	<b>6</b>
<b>6. Evaluation of Results.....</b>	<b>9</b>
<b>7. Refined Cleaning .....</b>	<b>9</b>
<b>8. Updated Results .....</b>	<b>16</b>

# 1. Project Description

## 1.1 Dataset Description

Restaurant inspections are conducted by the Food Protection Division of the Chicago Department of Public Health (CDPH). This ensures that food served to the public at licensed food establishments follows food safety guidelines. Inspections are performed on retail food establishments such as restaurants, grocery stores, hospitals, convenience stores, and schools. The inspections focus primarily on food handling practices, product temperatures, personal hygiene, facility maintenance, and pest control.

Data with reference to inspections that have taken place in the Chicago area have been compiled into the dataset 'Food\_Inspections.csv'. The data was collected by the City of Chicago Department of Health. The data includes inspection date, results, violations noted, business name, latitude and longitude location, license number, and risk. The dates covered are 01/02/2013 to 08/28/2017.

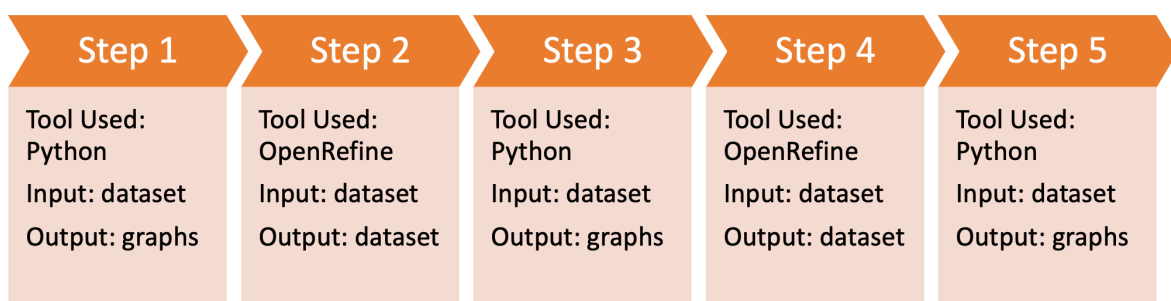
## 1.2 Use Case

- From the year 2010 to the year 2017, how did inspection results change for each facility type? For instance, which type of food provider keeps passing the inspection or keeps failing the inspection?

## 1.3 Target Columns

- Facility Type
- Results
- Inspection Date

# 2. Process Overview



The end goal of my research is to determine how inspection results for the various facility types changed from 2010 to 2017.

In Step 1, I imported the dirty data into a Python Jupyter notebook and explored the data in order to determine how to approach my research question. I also executed the analysis on the data. I performed some cleaning tasks, including transforming the 'Inspection Date' column to a datetime variable and extracting the year to create a 'Year' column in the corresponding dataframe.

Next, in Step 2 I imported the data into OpenRefine to perform the initial cleaning. This includes addressing the main data quality problems: non-uniform casing, typos, formatting, and blank entries. To address non-uniform casing, I performed a text transformation to force all entries into an uppercase format. To fix typos, I used several iterations of clustering and manual edits. The only formatting issue I had to fix was changing the `Inspection Date` column from a text variable to a date variable. At this stage, I decided to leave all of the blank entries in order to preserve the distribution of the data.

In Step 3, I used the same Jupyter notebook to import the clean data and execute the analysis on the data.

In Step 4, I used OpenRefine to continue cleaning the data. In this iteration of cleaning, I used GREL expressions to predict the `Facility Type` based on the `DBA Name` of a facility. I also condensed the number of Facility Types by manually determining the overarching category of each Facility Type.

Finally, in Step 5, I imported the clean data into my Jupyter notebook and executed the analysis on the data.

### 3. Execute on the Dirty Dataset

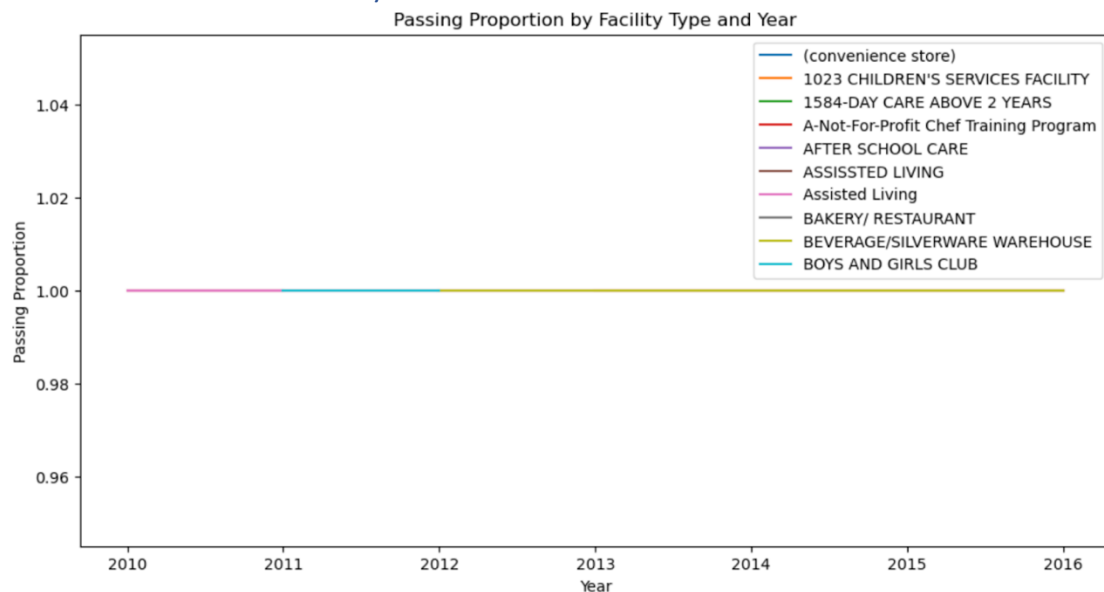
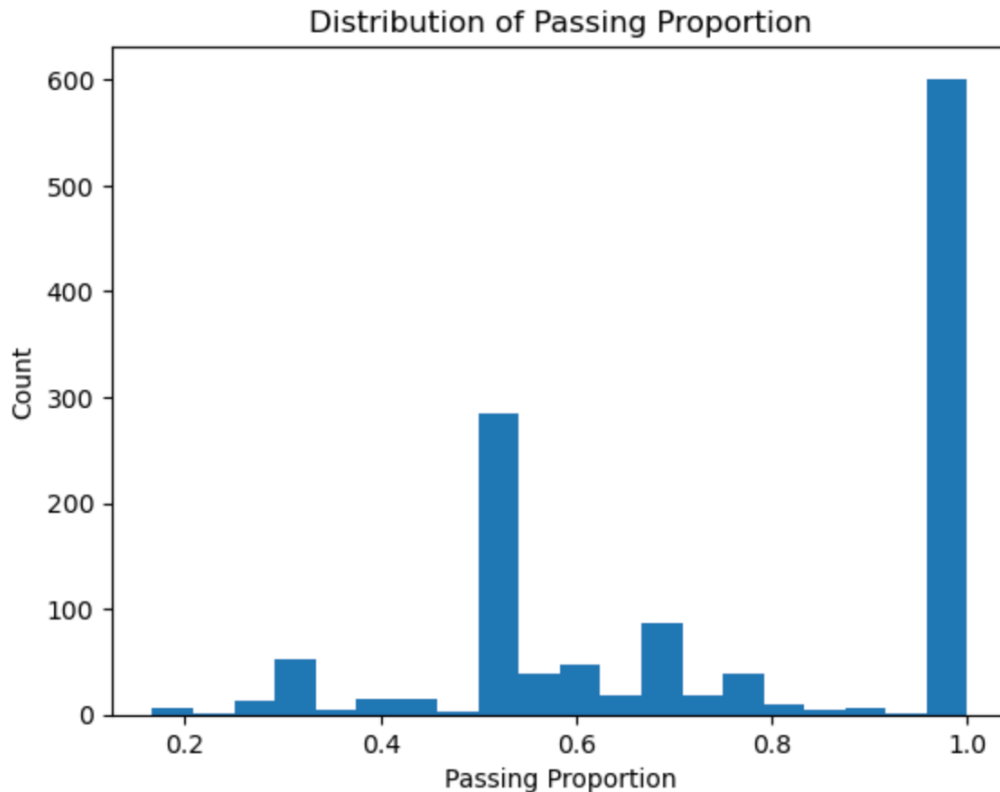


Figure 1.1



*Figure 1.2*

In my initial analysis of the dirty data, I created a `Passing Proportion` column in which I calculated the percentage of passing facilities for each Facility Type. This was necessary for the exploration of the passing results of Facility Types over time. I created a line plot to examine the top ten Facility Types with the highest passing results, shown in Figure 1.1. These results are unconvincing, since they seem to be the first ten Facility Types that appear in the dataset when sorted in alphabetical order. I decided to explore the distribution of the `Passing Proportion` column. This visualization is shown in Figure 1.2. The figure shows that around 600 of the facilities have a passing proportion equal to 1. This demonstrates the necessity of clustering and further cleaning of the `Facility Type` column, as it's possible that many of the facilities have only a small number of facilities in their Facility Type categorization.

## 4. Data Errors and Quality Problems

### 4.1 Typos

1023 CHILDERN'S SERVICE  
FACILITY 8  
1023 CHILDERN'S SERVICE S  
FACILITY 8  
1023 CHILDERN'S SERVICES  
FACILITY 26  
1023 CHILDREN'S SERVICES  
FACILITY 3

*Figure 2.1*

### 4.2 Non-Uniform Casing

1023-CHILDREN'S SERVICES  
FACILITY 22  
1584-DAY CARE ABOVE 2  
YEARS 2  
A-Not-For-Profit Chef Training  
Program 5  
Adult Family Care Center 3

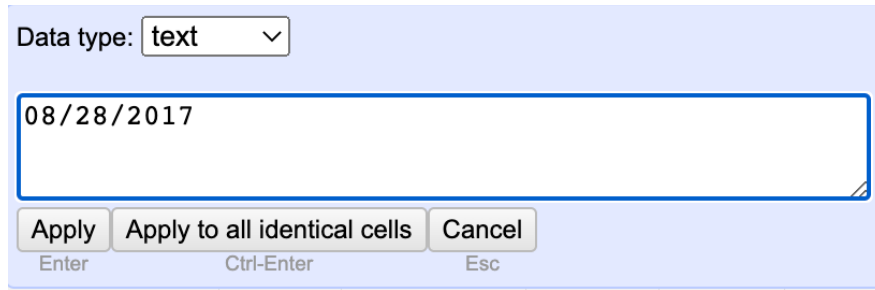
*Figure 2.2*

### 4.3 Blank Entries

(blank) 4560

*Figure 2.3*

### 4.4 Formatting



Data type: text ▼

08/28/2017

Apply Apply to all identical cells Cancel

Enter Ctrl-Enter Esc

Figure 2.4

Many of the data quality issues are present in the `Facility Type` column. These include typos, non-uniform casing, and blank entries. In terms of formatting, the `Inspection Date` column is in text format when it would be more useful in date format.

## 5. Initial Cleaning

<u>Column</u>	<u>Number of Changes</u>	<u>Transformations Applied</u>	<u>Related Data Quality Issues</u>
Facility Type	139 edits	<ul style="list-style-type: none"> <li>Text transform (toUppercase)</li> <li>Clustering</li> <li>Manual edits</li> </ul>	<ul style="list-style-type: none"> <li>Typos</li> <li>Non-uniform casing</li> </ul>
Inspection Date	1 edit	<ul style="list-style-type: none"> <li>Text transform (toDate)</li> </ul>	<ul style="list-style-type: none"> <li>Formatting</li> </ul>
Results	0 edits	None	None

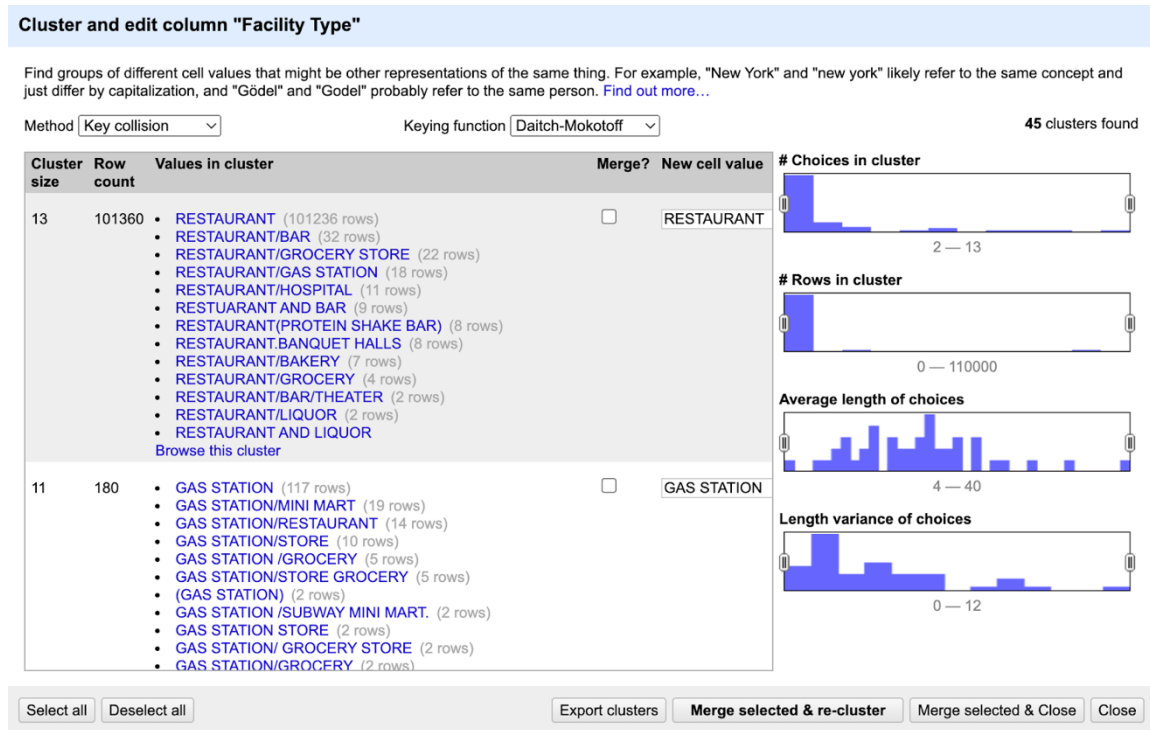


Figure 3.1

When performing the initial data cleaning, I first used the "To Uppercase" function in OpenRefine to transform the `Facility Type` column to all uppercase values. The choice of to Uppercase instead of to Titlecase or to Lowercase was arbitrary – I was simply looking for a way to make the format of the entries uniform.

Next, I used several iterations of clustering and manual typo correcting to fix the typos present in the `Facility Type` column. For my first iteration of clustering, I used Key Collision: Daitch-Mokotoff. The choice of clustering function was somewhat arbitrary, although I tried to pick the function that clustered the largest amount of cells. In this iteration, I did not choose all of the proposed clusters. As shown in Figure 3.1, one cluster included combinations of restaurants and other facilities. I decided to keep the distinction between these Facility Types. My second iteration of clustering used Nearest Neighbor: PPM. Again, I looked for the function that clustered the largest amount of cells. My third iteration of clustering used Nearest Neighbor: Levenshtein. In these second and third iterations, I manually checked each cluster and ended up selecting all of the clusters generated by the algorithm. I decided to combine all Daycare subsections (2-6 Years, Under 2 Years, 0-6, etc.) into one Daycare Facility Type. For the purposes of my research, I believe it is sufficient to look at one Daycare category.

Next, I used a text transform on the `Inspection Date` column in order to make the data easier to work with.

Finally, because there is a substantial amount of blank entries in the `Facility Type` column (as shown in Figure 2.3), I decided to leave them in the dataset.



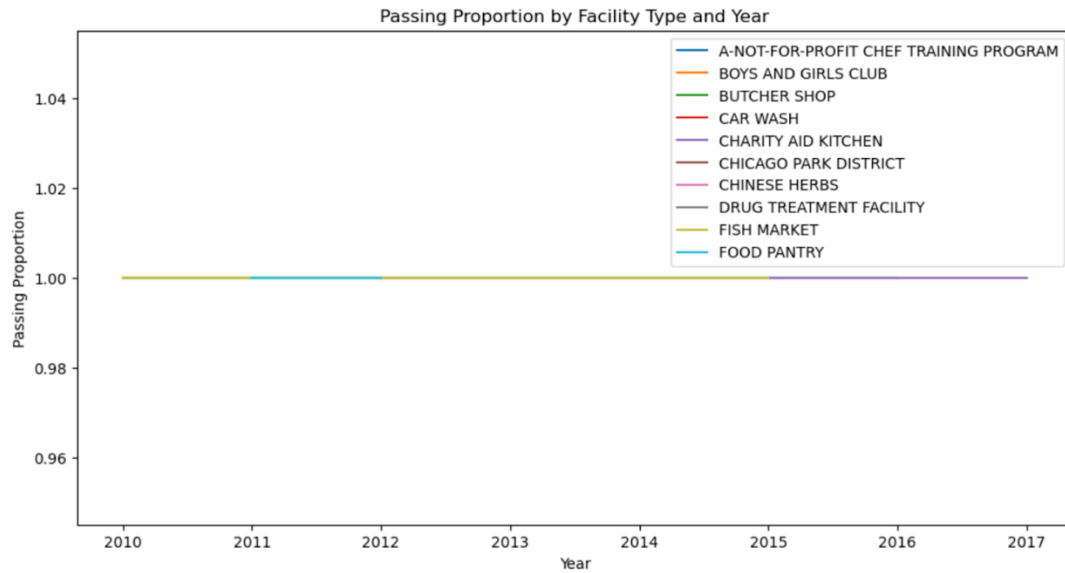


Figure 4.1

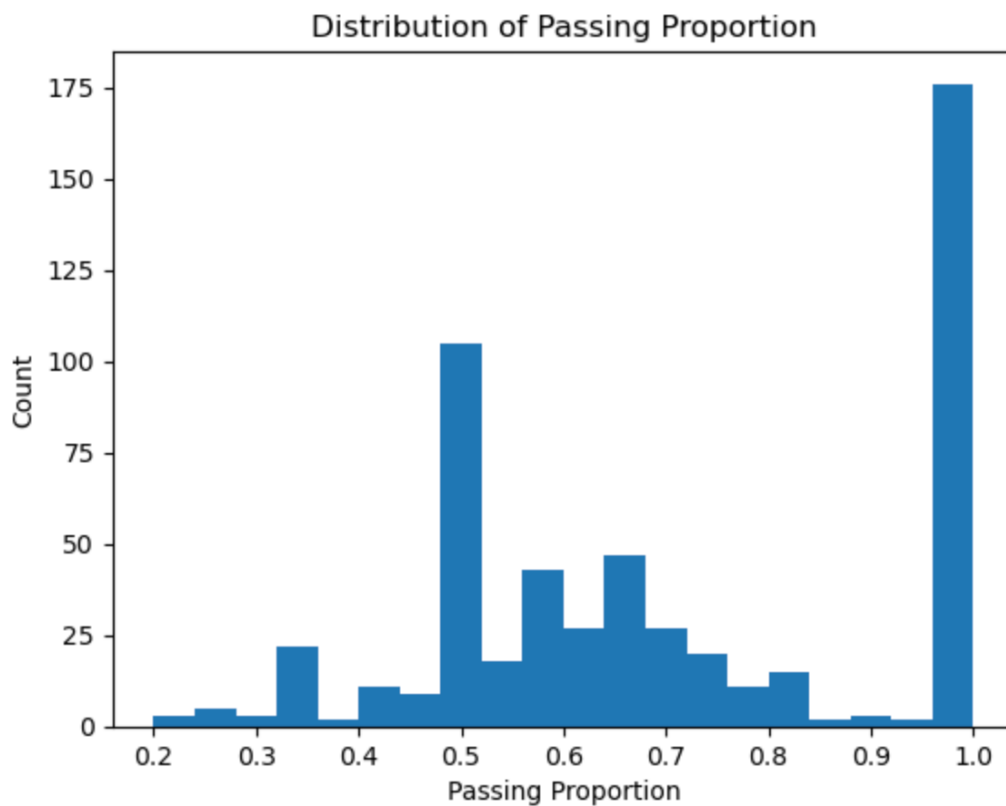


Figure 4.2

Again, I used Python to analyze the top ten facility types with the highest passing proportions. The analysis yields a different result, but the results still seem to contain the first

ten facilities in alphabetical order, shown in Figure 4.1. When looking at the distribution of the `Passing Proportion` in Figure 4.2, there is still a large amount of facilities with a passing proportion equal to 1 (around 175 facilities), but significantly fewer than were present in the dirty analysis.

## 6. Evaluation of Results

Although there were fewer facilities showing up in the group with passing proportions equal to 1 in the initial results, not much information can be obtained from the line plot in Figure 4.1. My goal is to get a better idea of how the passing proportion for various facility types changes over time.

It appears that there could be more data cleaning to be done. I thought that after cleaning the data, the results would differ more from those of the dirty dataset.

## 7. Refined Cleaning

One area in which I could improve the results of my data cleaning is the blank Facility Type values. My next steps would be to use GREL expressions to filter the blank expressions down to DBA Names that contain the word “restaurant”, “coffee”, “gas station”, etc. and change the Facility Type based on this word. The table below shows all of the GREL expressions performed.

<u>GREL Expression</u>	<u>Changed Facility Type to...</u>	<u>Number of Cells Changed</u>
value.contains(\"MART\")	GROCERY STORE	186
value.contains(\"RESTAURANT\")	RESTAURANT	321
value.contains(\"LIQUOR\")	LIQUOR	82
value.contains(\"CANDY\")	CANDY	12
value.contains(\"COFFEE\")	COFFEE	54
value.contains(\"GRILL\")	RESTAURANT	95
value.contains(\"PANTRY\")	PANTRY	55
value.contains(\"GROCERY\")	GROCERY STORE	99
value.contains(\"TAVERN\")	TAVERN	6
value.contains(\"CAFETERIA\")	CAFETERIA	6
value.contains(\"CAFE\")	CAFE	198

value.contains(\"BAKERY\")	BAKERY	52
value.contains(\"DUNKIN\")	COFFEE	28
value.contains(\"DINER\")	RESTAURANT	6
value.contains(\"MARKET\")	GROCERY STORE	114
value.contains(\"TAQUERIA\")	RESTAURANT	63
value.contains(\"SUBWAY\")	RESTAURANT	51
value.contains(\"DOLLAR\")	DOLLAR STORE	29
value.contains(\"CATERING\")	CATERING	40
value.contains(\"SUPERMERCADO\")	GROCERY STORE	17
value.contains(\"PIZZA\")	RESTAURANT	81

After performing these transformations, I was able to narrow down the number of blank Facility Type values from 4560 to 2965.

Another area in which I saw I could improve the data cleaning was in the number of Facility Types. I thought that many of the categories could be combined (e.g. the School category contains colleges and other schools). I manually looked at the Facility Type categories and condensed them down as I saw fit, using ChatGPT to provide suggestions. After 175 edits, I was able to condense the categories down to 19 different Facility Types. Below is the list of remaining Facility Types and the categories I grouped under each.

- BAKERY
- BAR/TAVERN
  - BAR
  - BREWERY
  - LOUNGE
  - NIGHT CLUB
  - TAVERN
  - WINE BAR
- CAFÉ
  - CAFETERIA
  - COFFEE
  - JUICE BAR
  - SMOOTHIE BAR
  - TEA BREWING

- CANDY
- COMMUNITY SERVICES
  - AFTER SCHOOL PROGRAM
  - ART GALLERY
  - BOYS AND GIRLS CLUB
  - CHARITY AID KITCHEN
  - CHICAGO PARK DISTRICT
  - CHILDRENS SERVICES FACILITY
  - CHURCH
  - FARMERS MARKET
  - FOOD PANTRY
  - MUSEUM/GALLERY
  - MUSIC VENUE
  - NON-PROFIT
  - RELIGIOUS
  - SHELTER
  - SOCIAL CLUB
  - SOUP KITCHEN
  - THEATER
  - VFW HALL
- DAYCARE
- GROCERY STORE
  - BUTCHER
  - BUTCHER SHOP
  - CHINESE HERBS
  - CONVENIENCE STORE
  - DRUG STORE
  - FISH MARKET
  - FOOD VENDING MACHINES
  - GAS STATION
  - GENERAL STORE
  - LIQUOR
  - MEAT MARKET
  - PANTRY
  - POULTRY
  - PRODUCE STAND
  - SNACK SHOP

- GYM/HEALTH CARE STORE
  - FITNESS CENTER
  - GYM
  - GYM STORE
  - HEALTH CARE STORE
  - HEALTH CENTER
  - HERBAL STORE
  - NUTRITION STORE
  - PROTEIN SHAKE BAR
- HOSPITAL
- HOTEL
  - ROOM SERVICE
- ICE CREAM
  - FROZEN DESSERT PUSHCARTS
  - GELATO SHOP
  - MOBILE FROZEN DESSERTS
  - PALETERIA
- KIOSK
- OTHER
  - AIRPORT LOUNGE
  - ANIMAL SHELTER CAFÉ PERMIT
  - CAR WASH
  - COLD/FROZEN FOOD STORAGE
  - COMMISSARY
  - DISTRIBUTION CENTER
  - DRUG TREATMENT FACILITY
  - EMPLOYEE KITCHEN
  - HOOKA LOUNGE
  - ILLEGAL VENDOR
  - INCUBATOR
  - KITCHEN
  - KITCHEN DEMO
  - LAUNDROMAT
  - LIMITED BUSINESS
  - MASSAGE BAR
  - MEAT PACKING
  - MOBILE FOOD

- MOBILE PREPARED FOOD VENDOR
- NAIL SHOP
- NEWSSTAND
- NORTHERLY ISLAND
- PACKAGED FOOD DISTRIBUTION
- PEDDLER
- POOL
- PRE PACKAGED
- PUSH CARTS
- REGULATED BUSINESS
- REHAB CENTER
- REPACKAGING PLANT
- RESEARCH KITCHEN
- RIVERWALK
- SHARED KITCHEN
- SPA
- STADIUM
- SUMMER FEEDING
- TRUCK
- UNLICENSED FACILITY
- UNUSED STORAGE
- URBAN FARM
- VENDING MACHINE
- WAREHOUSE
- WATERMELON HOUSE
- WEIGHT LOSS PROGRAM
- WHOLESALE
- RESTAURANT
  - DELI
  - DONUT SHOP
  - HOT DOG CART
  - ROOFTOP
  - SMOKEHOUSE
  - TENT RSTAURANT
- RETAIL
  - BLOCKBUSTER VIDEO
  - BOOKSTORE

- CELL PHONE STORE
- DOLLAR STORE
- PHARMACY
- POPCORN SHOP
- STORE
- TEA STORE
- VIDEO STORE
- WINE STORE
- SCHOOL
  - A-NOT-FOR-PROFIT CHEF TRAINING PROGRAM
  - ALTERNATIVE SCHOOL
  - CHARTER SCHOOL
  - COLLEGE
  - CULINARY SCHOOL
  - DINING HALL
  - PASTRY SCHOOL
  - TEACHING SCHOOL
- SENIOR CARE FACILITY
  - ADULT FAMILY CARE CENTER
  - ASSISTED LIVING
  - LONG TERM CARE
  - NURSING HOME
  - SENIOR DAY CARE
  - SUPPORTIVE LIVING FACILITY
- SPECIAL EVENTS
  - BANQUET
  - CATERING
  -
- (blanks)

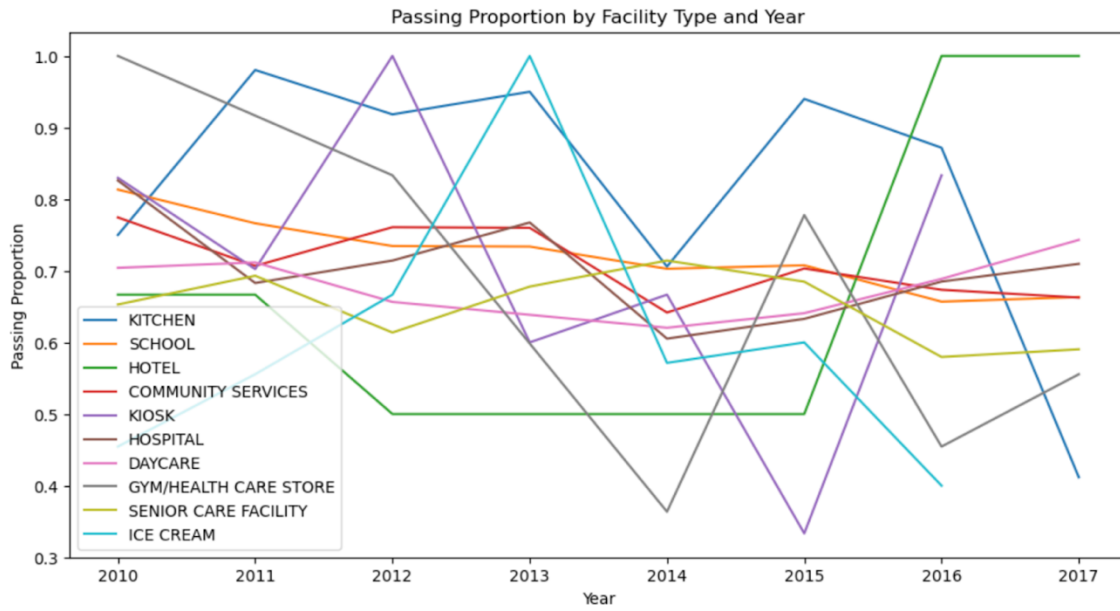


Figure 5.1

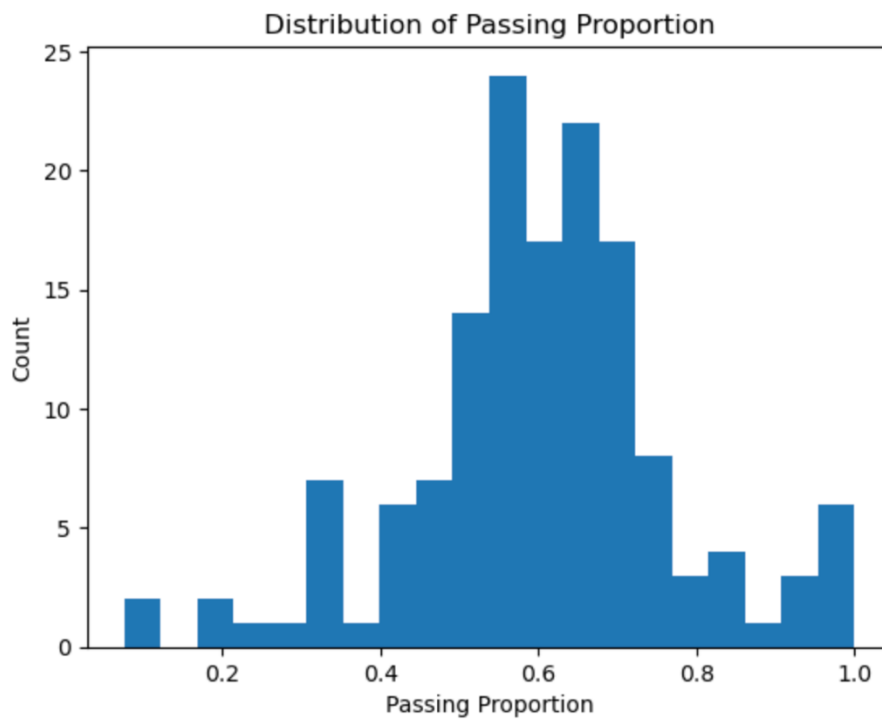


Figure 5.2



As shown in Figure 5.2, the distribution of passing proportion appears to follow more of a Normal distribution than the other two distributions. Based on this distribution, the changes in proportions should vary more over time. Figure 5.1 shows that the passing proportions for the top ten facility types with the highest passing proportions vary significantly over time. In addition, the list of the top ten Facility Types with the highest passing proportions is different from that shown in the other two line plots.

Much information about the top ten Facility Types can be extracted from Figure 5.1. Specifically, I found it interesting that for the Hotel Category, the passing proportion started in the middle range before decreasing for a while, and then working its way up to 1.

## 8. Updated Results

While not much information was obtained from the first two iterations of analysis, the refined cleaning had a large impact on my analysis. I gained numerous insights into the trends of passing proportions of facility types over time.

I believe I cleaned the data to the best of my abilities and there is no further cleaning to be done on my end. I was able to clean the dataset enough to gain information about trends in the passing proportion of various facility types. Given more time, I could possibly look further into the remaining blank values in the `Facility Type` column.