CHILDCARE CENTERS SAFETY LEVEL PREDICTION

Figure 1 : Picture is from https://www1.nyc.gov/site/doh/services/child-care.page

Introduction

What are the priorities of parents while looking for childcare for their beloved ones? Safety, hygiene, curriculum, and tuition are often at the top of the list and judging the items in the list is a tedious task. Typically, parents assess publicly available information; for example, they compare the tuition and curriculum of various childcare providers and decide. There are multiple sources to access school information, such as "Great School," “Yelp”, “Google”, or the business webpages of the specific institutions. Although the majority of these sources are crowdsourcing environments, the design elements of the presentation are market-driven; thus, critical topics such as safety and hygiene may go unnoticed. Specifically, for the early childhood education centers, the merits to judge the quality of the institution could be limited to the tuition and reviews of the parents. Alternatively, the quality assessment of childcare centers, or pre-kindergartens can be enhanced by using a domain-independent figure of merits, such as hygiene and safety violations. The data for these violations is publicly available; however, it requires systemic analytical processing so that it could become a means of decision support.

State and local governments are responsible for issuing license for child care centers. This responsibility includes periodic and follow-up inspections to the educational institutions, which operate in the jurisdiction district of the government. The results of inspections can be viewed online, or interested parents can request the results from the responsible agency. A typical inspection report includes some critical information regarding the violations and complaints that a childcare program has experienced. The problem is not only deciphering the report but also to communicate the priorities of the inspector to concerned parents.

Problem Statement

New York State has regulations that govern the minimum requirements for licensed and registered childcare centers in the state of New York. These requirements have been developed to provide a healthy and safe environment for the children. However, according to the National Institutes of Health, just 10% of childcare centers provide high-quality care. There are 1,789,069 children under 18 years old in New York City. Raising a child for many families -in such an expensive city- necessitates both parents working full time. One of the biggest concerns of a working parent is to provide high-quality childcare for her/his children. There are several methods to utilize when seeking a reliable childcare center by parents. Looking for online reviews, trying to access the most recent inspection reports from the state, and finding a neighbor or a friend who have suggestions are few of them. Even though all these methods support the decision process; the entire process could be exhausting and may not efficient for many parents. This capstone project investigates the factors that affect the safety level of childcare centers; it provides a systemic analytical analysis of the data that can entail decision support for concerned parents.

Clients

Parents and childcare services are possible stakeholders of this project. Also, it can be used by private, and public schools, community centers, and religious institutions for better service optimization & maximize the safety and hygiene for their students.

Data Set

1. The New York City Childcare Center Inspections data can be found on DATA.GOV: DOHMH Childcare Center Inspections

The DOHMH(Department of Health and Mental Hygiene) Child Care Inspection Data is formed by New York City childcare inspection office. The data includes preschools, daycare centers, school-based childcare centers, and camps located in New York City. It presents the childcare program's inspection history over the past three years (between 2016 and 2019). Each row represents the summary of an inspection visit to a specific childcare center. There are three types of inspection visits, regular, compliances, and based on reports. A regular visit is for a general investigation. A compliance visit is to follow up with the facility for the violations, which are detected at the previous visit. The last type of visit is for reported facilities.

According to data, 'Violation Rate Percentages' is a crucial parameter to understand the problems in facilities. Nevertheless, it is hard to comprehend how inspectors calculate the value of the violation rate. The only thing, which explained in the data is that the violation rate consists of 'Critical Violations,' 'Public Health Hazard Violations,' and 'General Violations.' The most serious type of violation is called 'Public Health Hazard Violations.' According to the New York City regulations, these violations must be corrected within one business day. According to the New York City regulations,' Critical Violations' are the second serious type of violations and must be corrected within two weeks. According to the New York City regulations, the general violations are minor and must be corrected within

one month. The New York City Health Department policies indicate that if the number of general violations is less than six, a compliance inspection is not required.

The variables used in this study are:

- **Center Name:** This is the name of the childcare center as known to the public.
- **Borough:** There are 5 main regions that the data collected; Bronx, Manhattan, Queens, Staten Island, and Brooklyn.
- **Zip Code:** Zip code as per the address of the entity.
- **Age Range**: the possible age range of children in the program.
- **Maximum Capacity:** The maximum number of children the facility is allowed, based on the square footage of class and playrooms, the number of toilets and sinks, and overall estimates from the NYC Department of Buildings. Enrollment can be higher than the maximum capacity if there are part-time programs.
- **Facility Type:** There are three types of facility in the data set; School Base Child Care (SBCC), Camp, Group Day Care (GDC).
- **Violation Rate Percent**: Percent of Initial Inspections that resulted in at least one Critical or Public Health Hazard violation. There is a huge lack of information in the coding of the data; as a result, terms pertaining to data, - such as “violation rate” – becomes difficult to decipher. To solve this problem number of violations are used as a predictive variable.
- **Total Educational Workers:** Current number of Educational Staff in the program, including teachers, assistant teachers, teacher directors and education directors.
- **Public Health Hazard Violation Rate:** Percent of Public Health Hazard violations among all violations issued at initial inspections during the past 3 years. If the same violation is cited multiple times during one inspection, it is counted only once.
- **Critical Violation Rate:** Percent of Critical violations among all violations issued at initial inspections during the past 3 years. If the same violation is cited multiple times during one inspection, it is counted only once.
- **Regulation Summary:** Violation Description Language or null/blank for no violation in the inspection.

- **Violation Category:** There are four type of entries for violation categories; Public Health Hazard (violation needs to be fixed within 1 day), Critical (Violation needs to be fixed within 2 weeks), General (Violation needs to be fixed in 30 days) or null/blank for no violation in the inspection.
- **Violation Status**: Violation is either open, or corrected, or no violation.
- **Inspection Summary Result:** The summary result of the inspection, based on the type of inspection and the number and types of violations.

2. The second data set is retrieved from yelp.com with a required API. The data set used for latitude, and longitude of childcare centers, and the ratings.

3. The third data set is used to visualize the data and the zip code boundaries on the map. The shape file is from JSSPINA.CARTO:

New York City zip code tabulation areas polygons

Data Wrangling

1. **Unnecessary information columns:** At this stage, I filtered out unnecessary columns and dropped them.

```
df=df(columns=['column_name'])
```

2. **Duplicated rows:** Duplicates are data points which are repeated rows in the dataset. I removed the duplicated rows by use of 'drop_duplicates()' method.
3. **Missing data:** I handled missing data in two steps. First of all, I kept only the rows with at least 23 non-NA values.

```
df=df.dropna(thresh=23)
```

Then, I used 'fillna' method to replace the missing values with zero.

```
df=df.fillna(0)
```

4. **Incompatible types of columns:** Incompatible types of columns: The variables of 'ZipCode' column were not compatible with the column type. Therefore, I added an argument to the importing the data code to change it from 'float' to 'string.'

```
data=pd.read_csv('DOHMH_Childcare_Center_Inspections.csv',
dtype={'ZipCode': 'str'})
```

5. **White spaces in column names:** There were some spaces between the word of column names. I removed these white spaces.

```
df.columns = df.columns.str.replace(' ', '')
```

6. **Bad Data Problem:** There was a bad data problem in ‘Facility Type’ column. The Camp variable used to exist as ‘Camp’ and ‘CAMP’. I solved the issue with the following code:

```
data_c.FacilityType=data_c.FacilityType.str.lower()
```

7. **A New Column:** I created 'Operated_time' column through subtracting two columns.

```
data_c['Operated_time']=data_c['PermitExpiration']-data_c['DatePermitted']
```

Before the subtraction, I changed the type of these columns from object to datetime.

```
data_c['PermitExpiration']=data_c['PermitExpiration'].apply(pd.to_datetime)
```

I also changed the 'Inspection Date' columns' type object to datetime.

Key Findings

During exploratory data analysis, we ask the following questions:

1. What kind of factors affects the safety level of childcare facilities?
2. Is there a relationship between the inspection report findings and the ratings of the parents?
3. Does location/neighborhood affect the safety level of childcare centers?

Initial findings are:

- As we expected, the number of schools are correlated with the number of violations. Brooklyn is the region, which has the highest number of schools in the New York City area, and Staten Island has the least number of schools. The highest number of violation has been

observed in the Brooklyn region and the number of observed violations in Staten Island is the smallest.

- The number of group daycare is higher than other types of facilities in each region.
- 3698 inspection visits turn out without any violation observation.

· The number of inspections visit that violation observed is 38081.

- The maximum number of inspection visit to a facility is 74. It means approximately 24 visit in a year and 2 visit in a month.
- There isn't a correlation between the number of educated staff of a facility and the number of the violation. However, many parents trust facilities more if educated staff number is higher.
- There is a relation between the number of years that the facility operated and the number of the violation. The longer time operated facilities have higher violation numbers.
- There isn’t any educated staff in camps.
- The highest number of violations observed between 2-5 year-old age ranges.
- The number of violation is least in summer camps. However, summer camps are short term facilities. They operate for three months. It could affect the amount of visit for inspection.
- There isn't a correlation between the number of violation and the ratings at Yelp. It means parents are not a good detector for violations.
- There are some neighborhoods that have facilities only with high number of violations.