Can a parent predict the safety level of a childcare center?

Predictive Analysis of Safety Level of New York City Childcare Centers

Daphne Canan / September 2019

Recent Incidents

- 3-month-old baby died in SoHo Daycare in July 2015
- SoHo Daycare is operated for 14 years without a license.
- 3-year-old boy died in a daycare in Manhattan on December 2018
- Elijah was given a grilled cheese sandwich at daycare, despite the staff being told he had a severe dairy allergy
- a 4-month-old baby died in a daycare in Brooklyn on January 2019

Data

- Health departments conduct inspections
- Department of Health and Mental Hygiene) Child Care Inspection Data
- Social media data imported from Yelp
- Each row represents the summary of an inspection visit
- Inspection history over the past three years
- Yelp data set used for latitude, and longitude of childcare centers, and the ratings

Features

- **Numerical Features:**

Maximum Capacity

Violation Rate Percent

Total Educational Workers

Public Health Hazard Violation Rate

Critical Violation Rate

- **Text Features:**

Regulation Summary

Inspection Summary Result

- **Categorical Features:**

Center Name

Borough

Zip Code

Age Range

Facility Type

Violation Category

Violation Status

Introduction

Data Wrangling

- Drop the unnecessary information columns
- Duplicated rows
- Missing data
- Incompatible types of columns
- White spaces in column names
- Bad Data Problem
- Creating New Columns

Exploratory Data Analysis

H0: There is a relation between the number of schools in a region and the number of the violation.

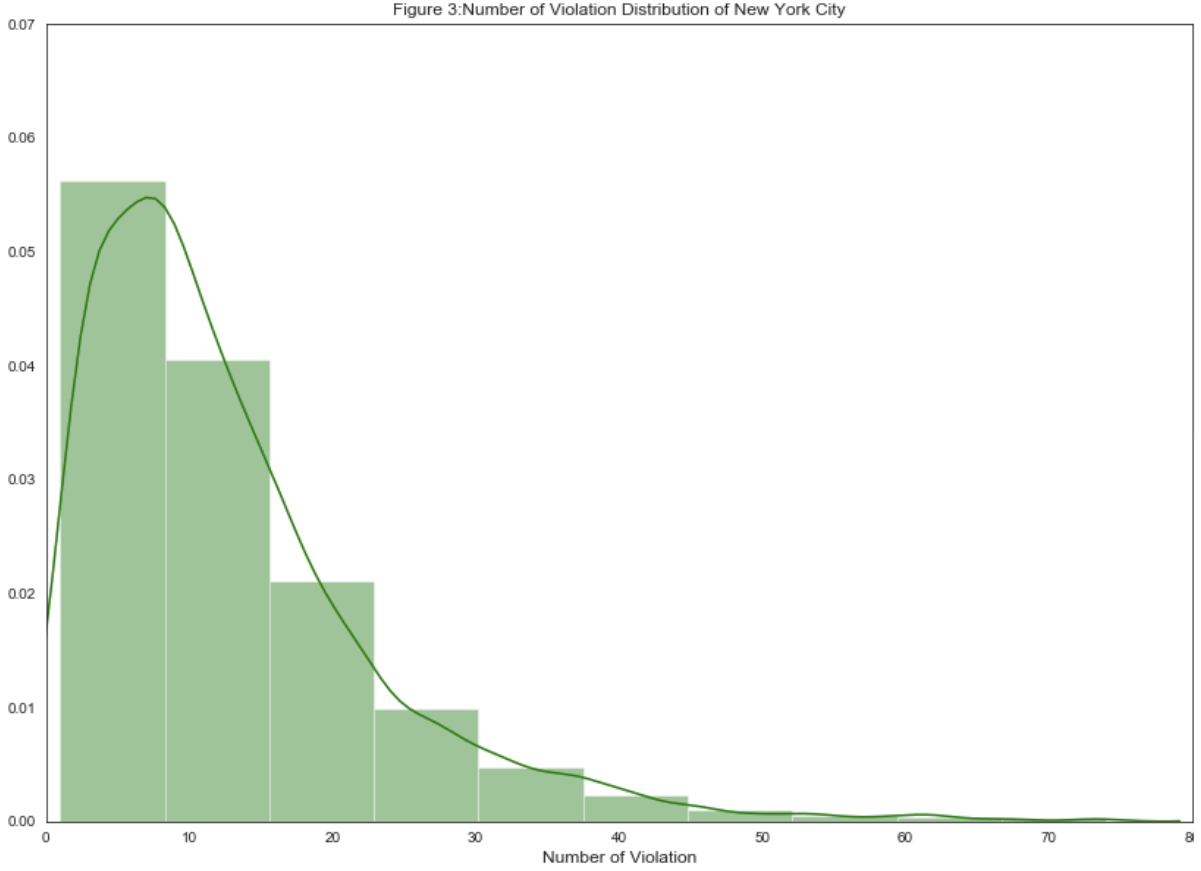
H1: There isn't a relation between the number of schools in a region and the number of the violation.

Table 1: Number of school vs Number of Violation Based on Regions

Regions	Number of Schools	Number of Violations
Bronx	438	7926
Manhattan	735	8319
Queens	681	8493
Staten Island	139	1676
Brooklyn	1219	11662

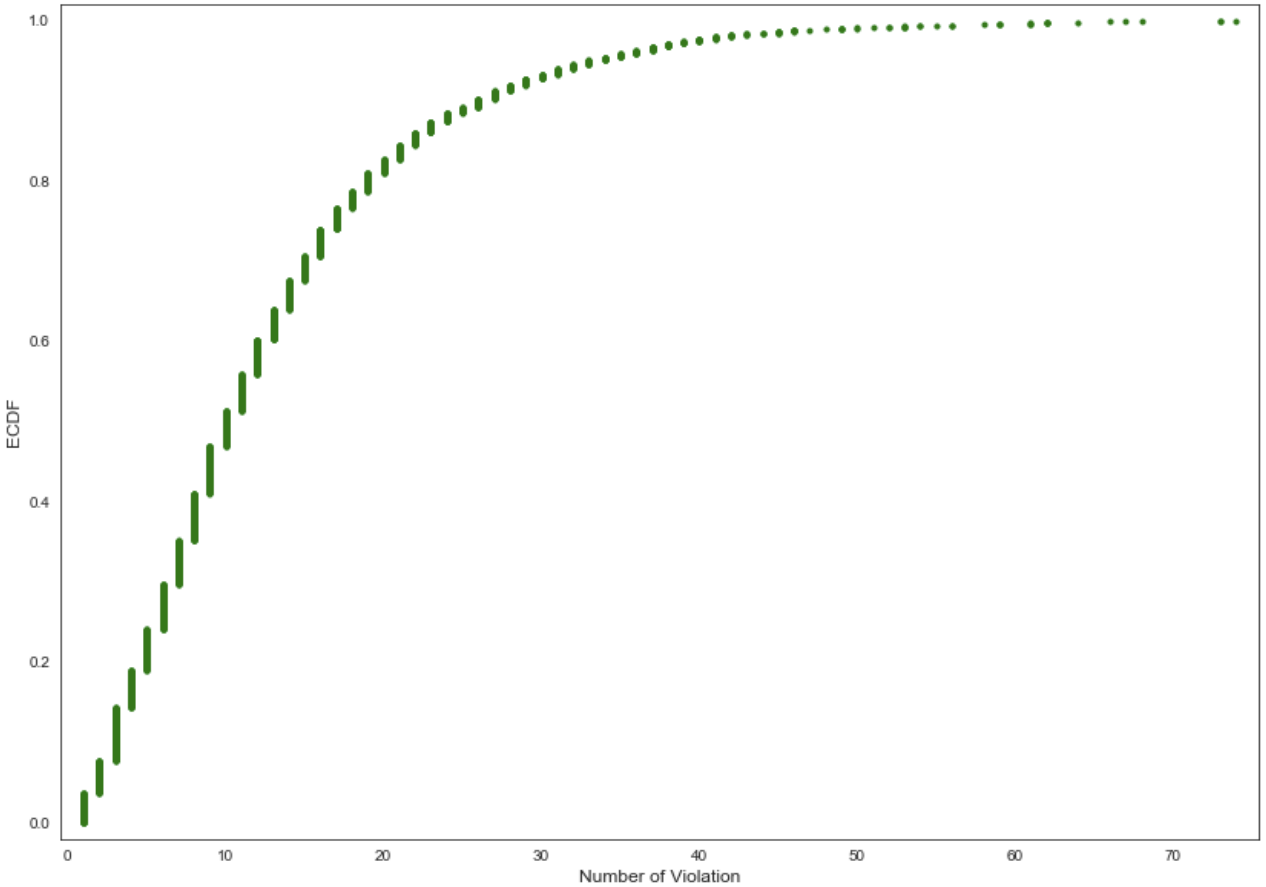
r: 0.921620192176706
p-value: 0.5128

Figure 3:Number of Violation Distribution of New York City



Exploratory Data Analysis

Figure 4:Empirical CDF Plot of New York City



Almost 80% of the schools have more than 20 violations

Key Findings

- 3698 inspection visits turn out without any violation observation.
- The number of inspection visit that violation observed is 38081
- The maximum number of inspection visit to a facility is 74
- The worst 3 child care centers are:
  - The Learning Tree
  - Tender Tots
  - Seabury Daycare Center

Figure 5:Empirical CDF Plot for Regions of New York City

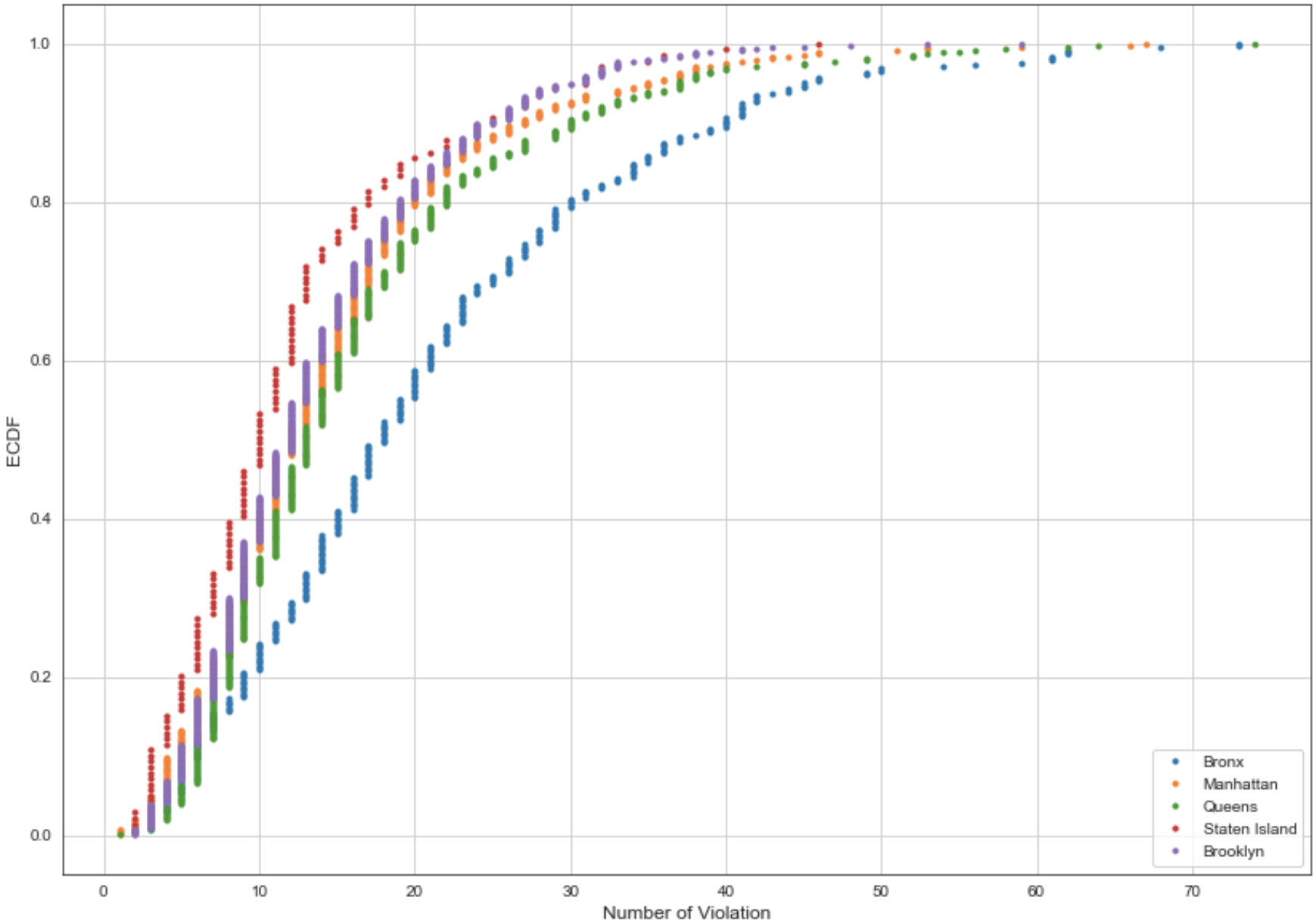


Figure 33: Relation Between Different Variables Based on the Facility Types

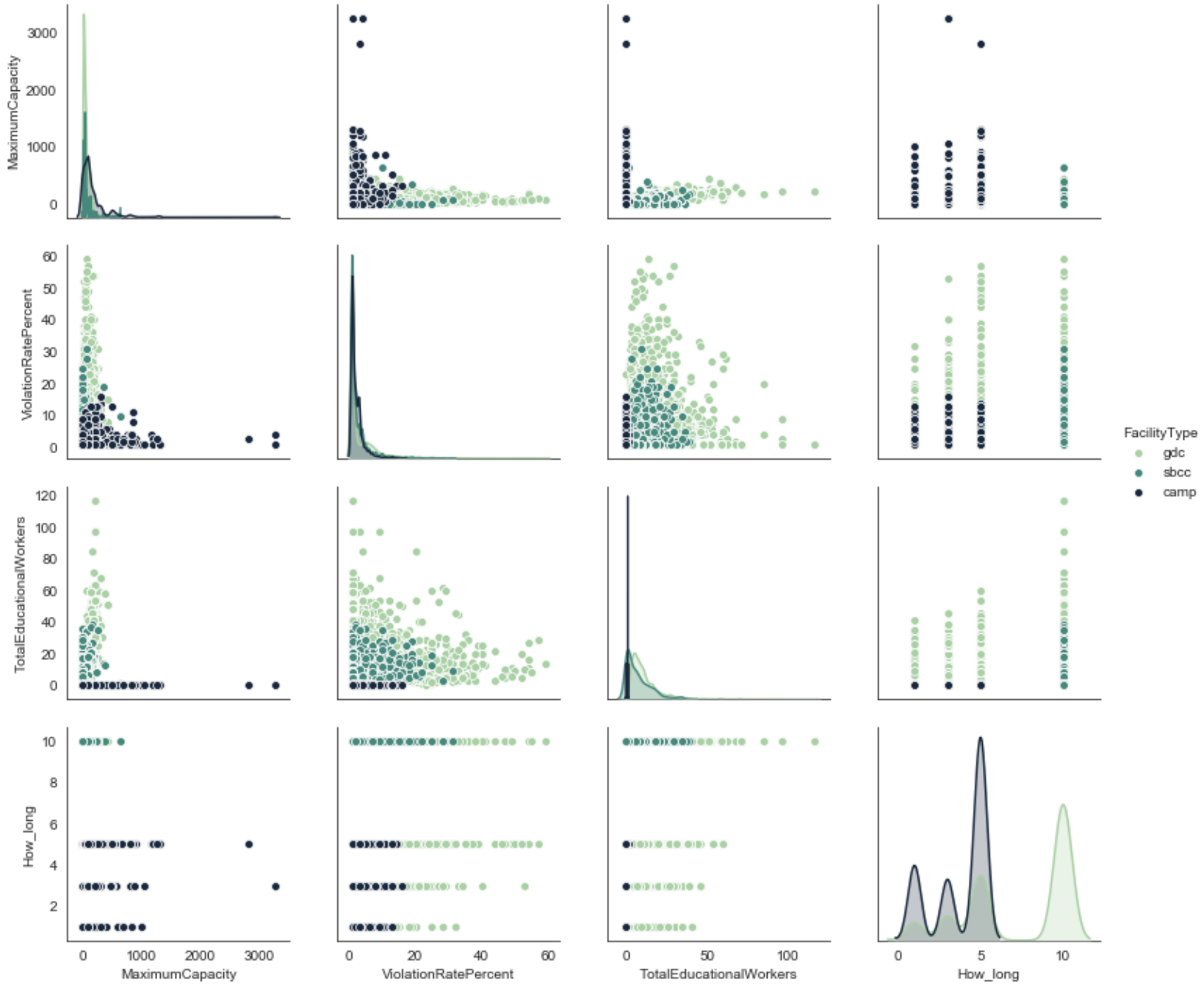
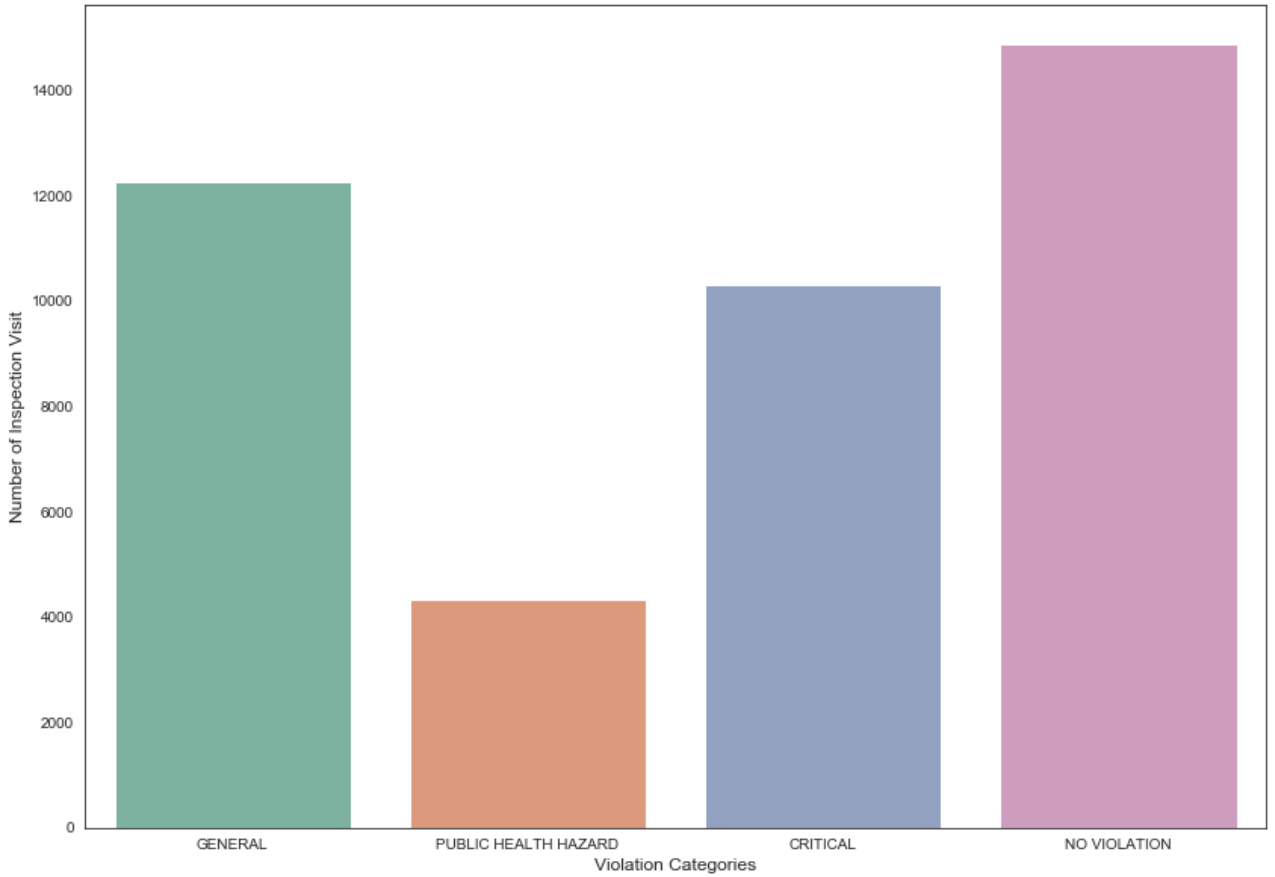
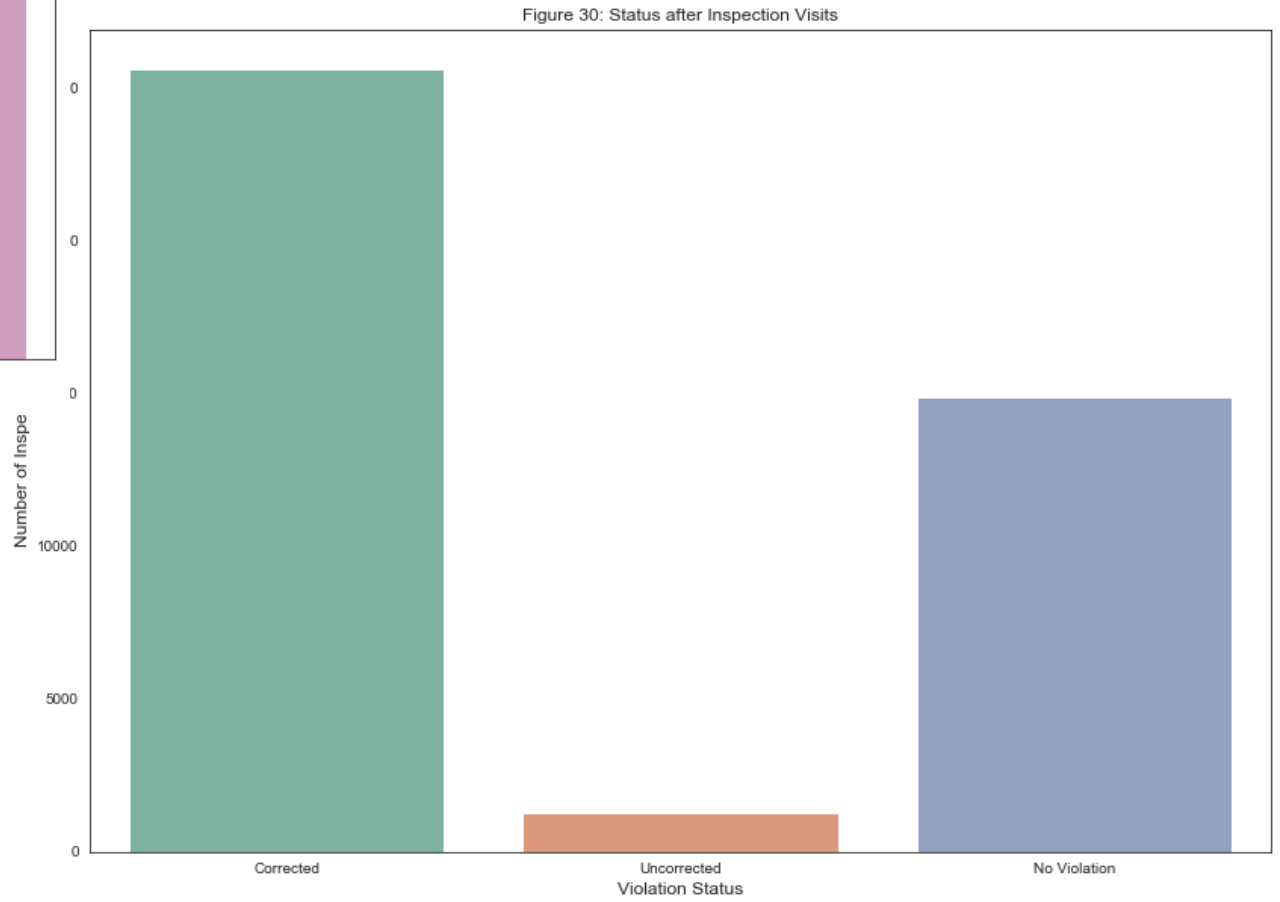


Figure 31: Number of Inspection Visit Based on Violation Categories



Exploratory Data Analysis

Figure 30: Status after Inspection Visits



- Number of uncorrected violation status is 1262
- Number of general violation is 12261
- Number of public health hazard violation is 4319
- Number of critical violation is 10318

Exploratory Data Analysis

How violation is understood?

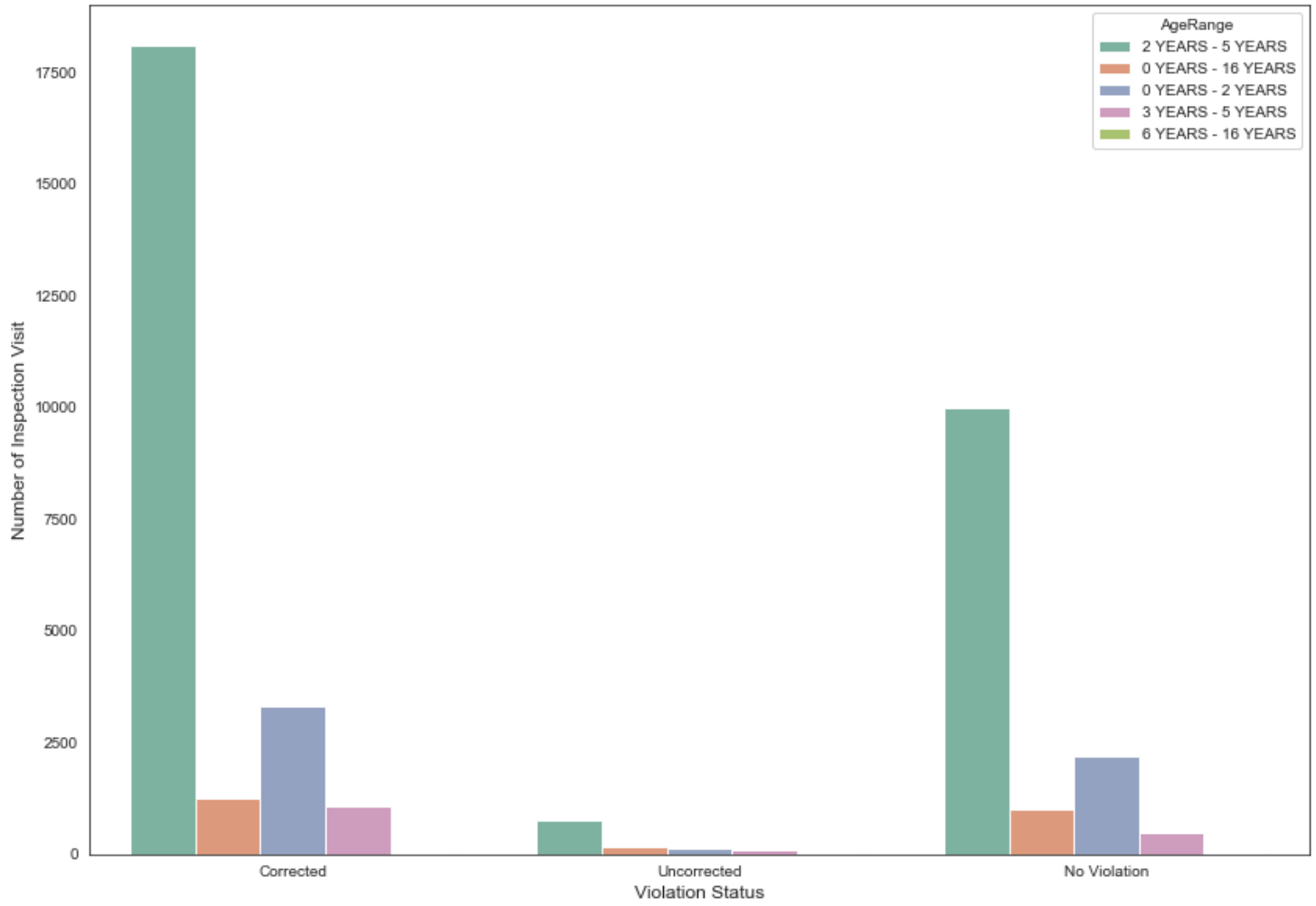
- Critical Violation Examples:
  - Pest problems
  - Allowing staff to perform when they are not healthy
  - Outdoor play area without fencing
  - Diaper changing area without proper sanitary
  - Fail to properly clean and sanitize plate and dishware
- Public Health Hazard Violation Examples:
  - Failed to stuff's criminal background checks
  - Employees failed to wash hands after using lavatory
  - Unsafe sleep environment for infants
  - Operating without a permit
- General Violation Examples:
  - Improper storage of garbage
  - Fail to obtain written parental permission for travelling
  - No isolation area for sick children
  - Smoking observed in outdoor area
  - Fail to provide adequate ventilation
  - Fail to provide clean sheets

Exploratory Data Analysis

The number of group daycare is the highest

The highest number of violations observed between 2-5 year-old age ranges.

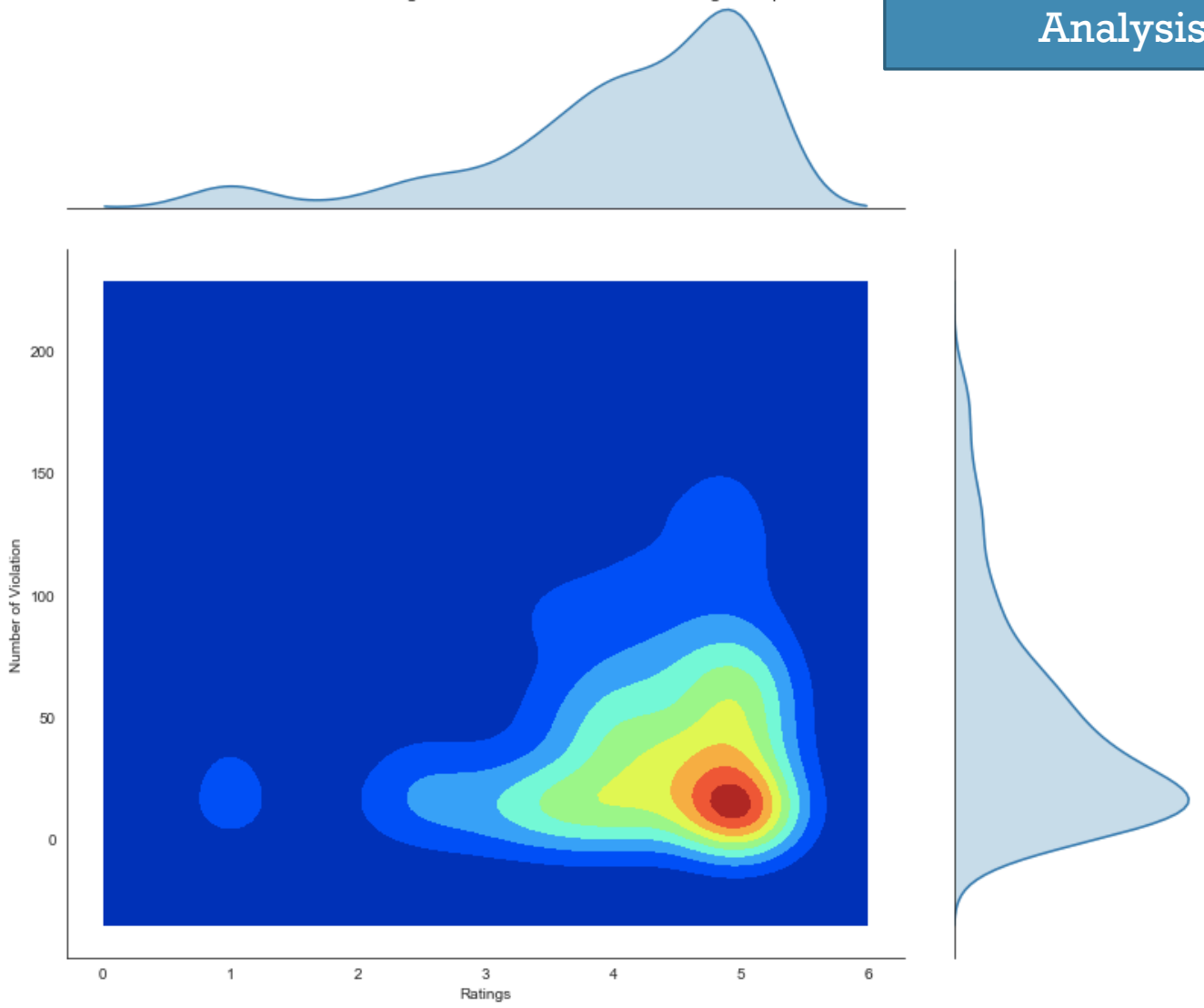
Figure 32: Violation Status based on Age Ranges



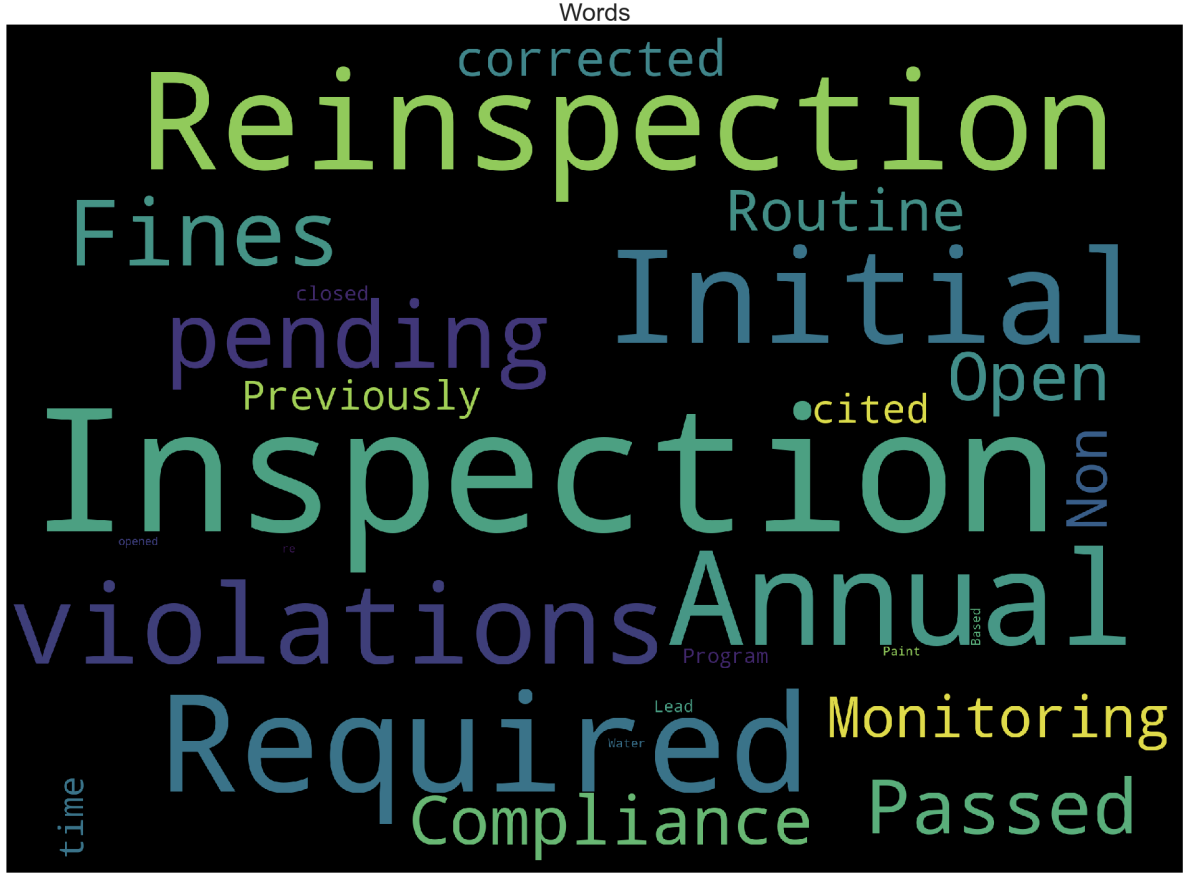
Exploratory Data Analysis

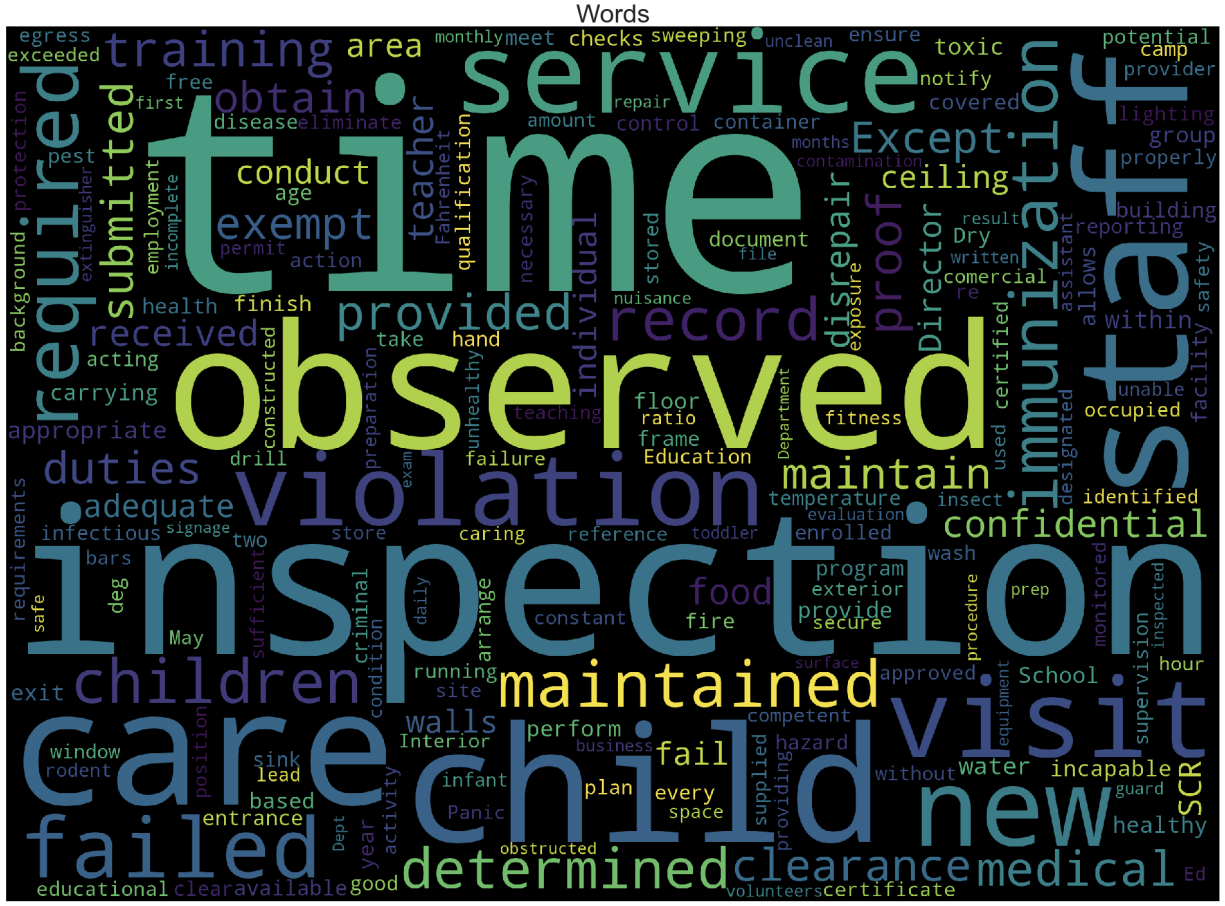
Parents are not accurate detectors for the safety level of a childcare center.

Figure 34: Number of Violation vs Rating at Yelp

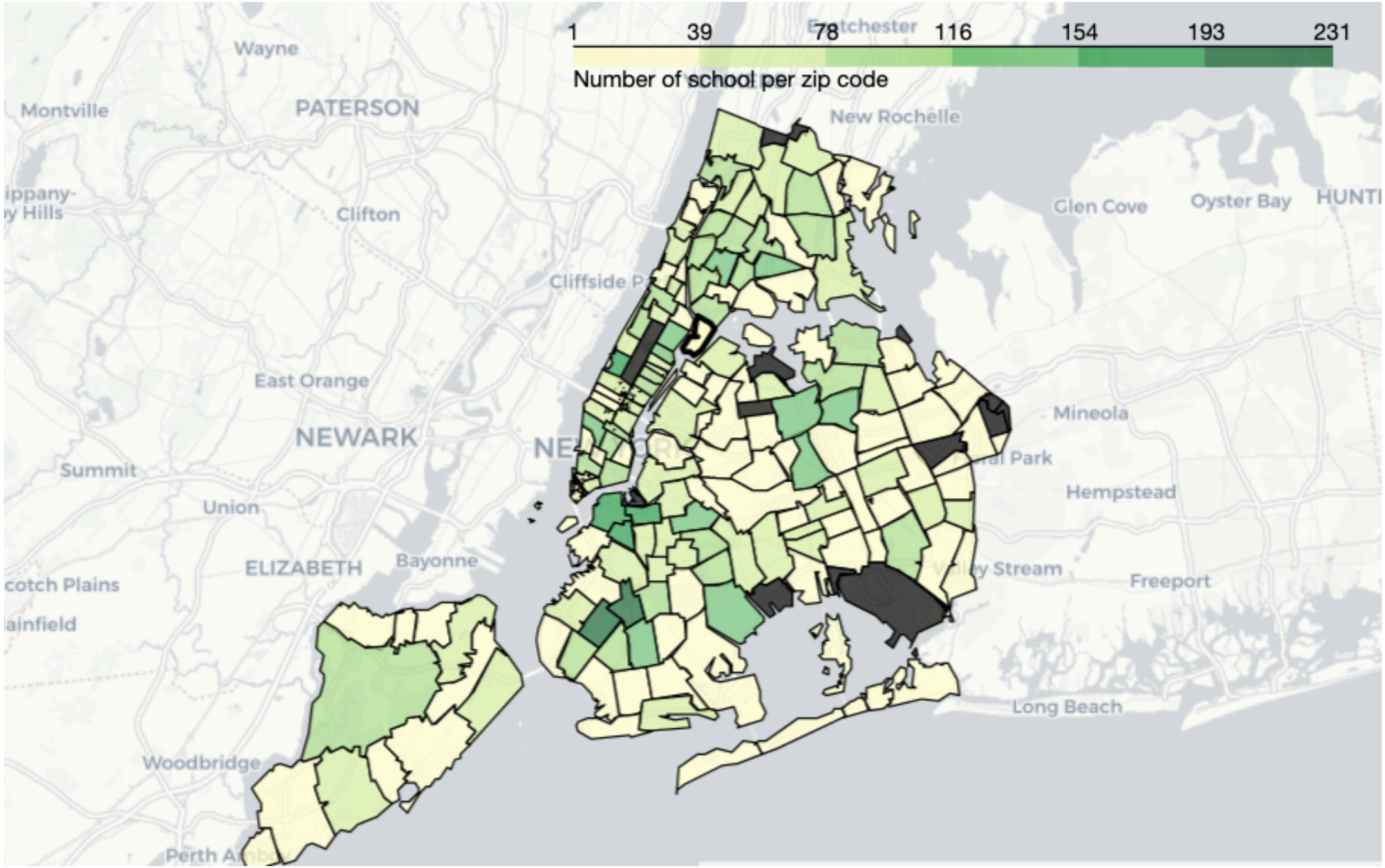


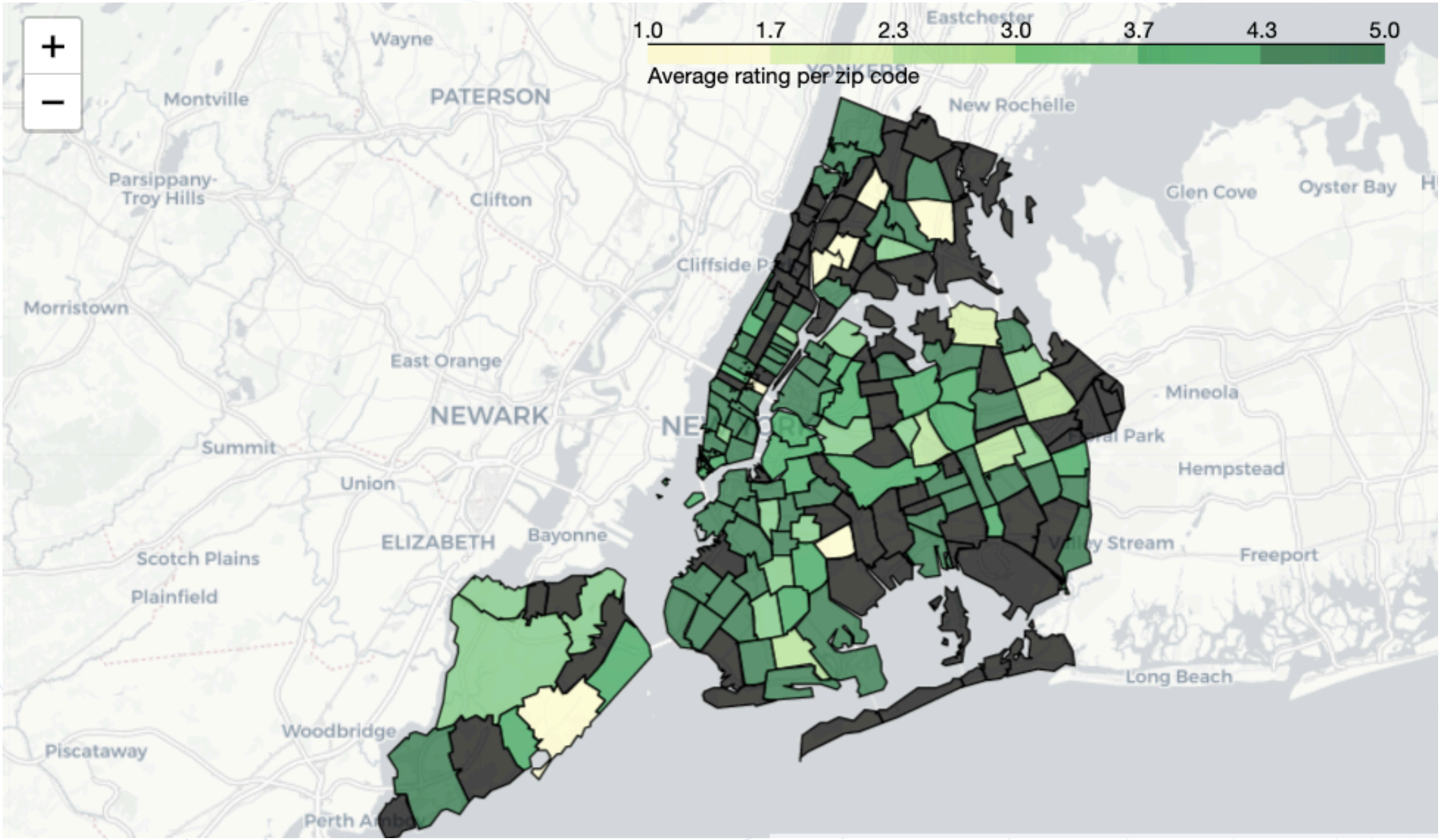
Exploratory Data Analysis





Word Clouds for Text Features


1
39
78
116
154
193
231
Number of school per zip code
Wayne
Montville
PATERSON
Clifton
East Orange
NEWARK
Summit
Union
ELIZABETH
Bayonne
Woodbridge
New Rochelle
Glen Cove
Oyster Bay
Mineola
Hempstead
Freeport
Long Beach


1.0
1.7
2.3
3.0
3.7
4.3
5.0
Average rating per zip code
Wayne
PATERSON
Montville
Parsippany-
Troy Hills
Clifton
Eastchester
New Rochelle
Glen Cove
Oyster Bay
Cliffside Park
Morristown
East Orange
NEWARK
Mineola
Summit
Union
Hempstead
ELIZABETH
Bayonne
Scotch Plains
Plainfield
Freeport
Long Beach
Woodbridge
Piscataway
Perth Amboy

Machine Learning Analysis

- Thresholds for safety levels:

value ≤ 4 ⇨ Safe

4 < value ≤ 12 ⇨ Warning

12 < value ⇨ Not Safe

- Number of values in each safety level
  - Safe : 1640
  - Warning: 11190
  - Not Safe: 28949
- Model for a multiclass and imbalanced data
- Up-sampling technique to remove class imbalance

Models
1 Random Forest Classifier (0.8907708593975556)
2 Gradient Boosting Classifier (0.8439996150514869)
3 Extra Tree Classifier (0.9764219035703975)

```
def create_pipeline():
    numerical_indices = [0,1,2,3,4,5]
    categorical_indices = [6,7,8,9]
    p1 = make_pipeline(PositionalSelector(categorical_indices),
StripString(), SimpleOneHotEncoder())
    p2 = make_pipeline(PositionalSelector(numerical_indices),
StandardScaler())
    feats = FeatureUnion([
        ('numericals', p1),
        ('categoricals', p2)
    ])
    pipeline = Pipeline([
        ('pre', feats),

('estimator',RandomForestClassifier(max_depth=15,n_estimators=100))
    ])
    return pipeline
```

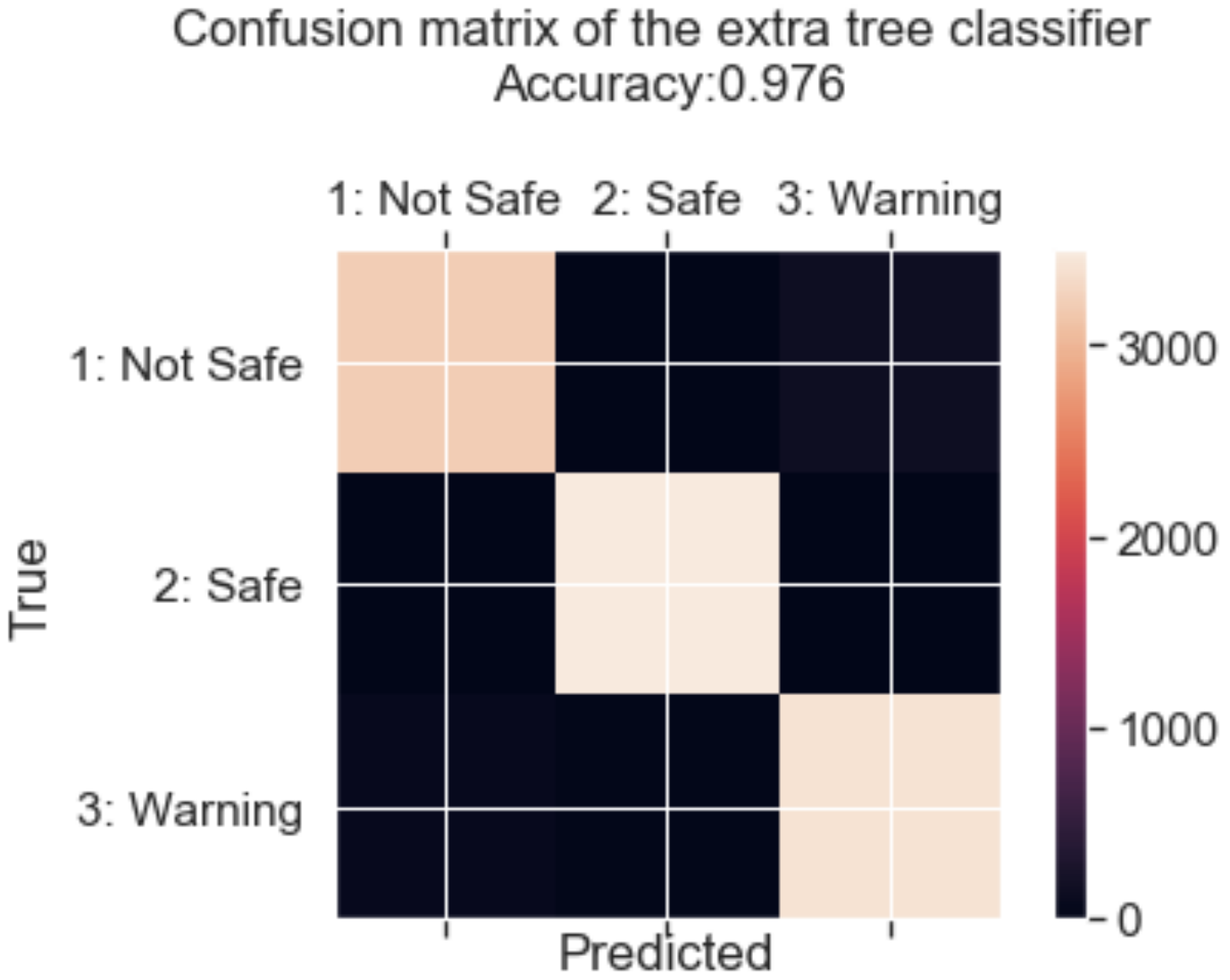
Extra Tree Classifier

Evaluation Metrics:

- Accuracy score: 0.9764219035703975
- Cross validation score: 0.9733210001607226
- Cohen Kappa score: 0.9646196951639845
- Hamming Loss: 0.0235780964296025 4
- Classification Report:

	precision	recall	f1-score	support
Not Safe	0.98	0.96	0.97	3363
Safe	0.99	1.00	1.00	3501
Warning	0.96	0.97	0.97	3527
micro avg	0.98	0.98	0.98	10391
macro avg	0.98	0.98	0.98	10391
weighted avg	0.98	0.98	0.98	10391

Confusion Matrix
[[[3215 1 147]
[0 3501 0]
[76. 21 3430]]



Feature Importance

- 'Violation Rate Percent', 0.0014042855508064466
- 'Inspection Summary Result', 0.0002917593830964625
- 'Public Health Hazard Violation Rate', 0.0008402926869307131
- 'Critical Violation Rate', 0.0001773861392154 1705
- 'Maximum Capacity', 0.00010761348261489092
- 'Total Educational Workers', 5.011412309221521e-06
- 'time', 8.981633317969614e-05
- 'Regulation Summary', 2.028875319022319 4e-06
- 'Borough', 1.7748047993655098e-05
- 'Facility Type', 1.1952568199891351e-05

Conclusion and Limitations

- No correlation between Yelp ratings and safety levels
- The risks that have been identified by inspectors
- Prioritizing which facilities may require more intensive inspections
- The system of assigning health code violations by inspectors may not be enough
- Lack of high participation of parents to rate the facilities
- Model not focused on a specific group of health code violations
- The prediction of violations can not forecast the behaviors of these facilities in the future.

Recommendations and Further Work

- Sharing the actual violation reason rather than violation scores may have practical impacts both for parents and childcare centers
- A system that parents can access easily, and quickly to communicate with authorities to get information and also share their reviews.
- Relation with the childcare center tuitions and the income level of regions
- Sentiment analysis of Yelp reviews