**Introduction**

Department of Health and Mental Hygiene (DOHMH) conducted inspections at active, city-regulated center-based child care programs and summer programs from July 2016 to May 2019 in New York City. The dataset contained any associated violations at these centers. According to DOHMH, these violations were pre-adjudicated, and the violations that were subject to potential fines were submitted to NYC office of Administrative Trials and Hearings to be formally judged as either sustained/upheld or dismissed.

This project reviewed and analyzed the dataset to help NYC families to make an informed decision in choosing child care centers based on the facilities’ program types and performance record with respect to past violations if any.

**Overview**

**Borough and Violation Rate Percent**

Violation rate percent represents a percent of initial inspections that resulted at least one critical or public health hazard violation. There are five boroughs in the dataset: Bronx, Brooklyn, Manhattan, Queens, Staten Island. The mean value of violation rate percent of five boroughs are not equal. Also, all the pairs of violation rate percent are statistically significantly different except for the pair State Island-Brooklyn and Staten Island -Manhattan. Bronx has higher violation rate percent than Brooklyn and it is statistically significant. Similar significances appear in all the pairs that included Bronx. The violation rate percent is affected by borough based on Pearson’s Chi-squared test. Violation rate percent is positively correlated with PHH violation rate and critical violation rate.

The results of statistical tests are presented in Appendix A.

**Borough and Public Health Hazard Violation Rate**

Public health hazard violation needs to be fixed within 1 day. The PHH violation rate represents the proportion among all violations that were issued at initial inspections in the last 3 years. The mean value of PHH violation rate of five boroughs are not equal. PHH violation rate is positively correlated with critical rate. Also, all the pairs of PHH violation rate are statistically significantly different at 0.99 or 99% confidence level. Lastly, PHH violation rate is dependent of the borough.

The results of statistical tests are presented in Appendix A.

**Borough and Critical Violation Rate**

Critical violation needs to be fixed within 2 weeks. The critical violation rate represents the proportion among all violations that are issued at initial inspections in the last 3 years.

The mean value of critical violation rate of five boroughs are not equal.

The results of statistical tests are presented in Appendix A. Lastly, critical violation rate is dependent of the borough.

**Facility Type and Violation Rate Percent**

Three facility types are presented in the dataset: Group Day Care (GDC – NYC DOHMH Permitted), SBCC (School Base Child Care – Licensed by NYS, DOE, Charter School or Regents), and Camp. The mean value of violation rate percent of the three facility types are not equal. Based on the statistical testing, there was no evidence that the violation rate percent is independent of facility types.

The results of statistical tests are presented in Appendix B.

**Program Type and Violation Rate Percent**

Four program types are presented in the dataset: Preschool, Infant Toddler, All Age Camp, and School Age Camp. Based on the statistical testing, there was no evidence that the violation rate percent is independent of facility types.

The results of statistical tests are presented in Appendix B.

**Maximum Capacity and Violation Rate Percent**

As the p-value is 0.9999, we do not reject the null hypothesis that the violation rate percent is independent of the maximum capacity. In other words, the violation rate percent is not affected by the maximum capacity of childcare centers.

The results of statistical tests are presented in Appendix C.

**Total Educational Workers and Violation Rate Percent**

Total educational workers represent the current number of educational staff in the program (teachers, assistant teachers, teacher directors, and education directors). The number is reported only upon inspection that it may not be up to date. Based on the dataset, the total educational workers ranged from 0 to 115, with mean value of 8.3 and median value of 6. There was almost no correlation between the total educational workers and the violation rate percent. However, another test indicates that the violation rate percent is not necessarily independent of the number of educational workers.

The results of statistical tests are presented in Appendix C.

**Appendix A.**

###correlation between viol\_rate, phh\_rate, crit\_rate

cor(viol\_rate, phh\_rate)

#0.6562685, p-value < 0.05

cor(viol\_rate, crit\_rate)

#0.8868788, p-value < 0.05

cor(phh\_rate, crit\_rate)

#0.4227864, p-value < 0.05

1. Borough and Violation Rate Percent
2. One-Way ANOVA Test

# Df Sum Sq Mean Sq F value Pr(>F)

#center\_table$borough 4 126579 31645 39.92 <2e-16 \*\*\*

# Residuals 2860 2266925 793

#---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Tukey’s Honestly Significant Differences

# Tukey multiple comparisons of means

# 99% family-wise confidence level

#

# Fit: aov(formula = center\_table$violation.rate.percent ~ center\_table$borough)

#

# $`center\_table$borough`

# diff lwr upr p adj

# BROOKLYN-BRONX -19.8426324 -25.2290866 -14.456178 0.0000000

# MANHATTAN-BRONX -18.6375109 -24.3473750 -12.927647 0.0000000

# QUEENS-BRONX -14.8061798 -20.6376423 -8.974717 0.0000000

# STATEN ISLAND-BRONX -19.7032752 -28.8071661 -10.599384 0.0000000

# MANHATTAN-BROOKLYN 1.2051215 -3.3204427 5.730686 0.9087772

# QUEENS-BROOKLYN 5.0364526 0.3584041 9.714501 0.0041930

# STATEN ISLAND-BROOKLYN 0.1393572 -8.2724805 8.551195 0.9999981

# QUEENS-MANHATTAN 3.8313311 -1.2157271 8.878389 0.0970627

# STATEN ISLAND-MANHATTAN -1.0657643 -9.6882739 7.556745 0.9944720

# STATEN ISLAND-QUEENS -4.8970954 -13.6006050 3.806414 0.3546301

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: tbl1

# X-squared = 400.72, df = 128, p-value < 2.2e-16

#######As the p-value is less than the .05 significance level, we reject the null hypothesis that ###the violation rate percent is independent of the borough.

1. Borough and Public Health Hazard Violation Rate
2. One-Way ANOVA Test

#Df Sum Sq Mean Sq F value Pr(>F)

#center\_table$borough 4 40774 10194 19.79 4.36e-16 \*\*\*

# Residuals 2860 1472983 515

#---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Tukey’s Honestly Significant Differences

# Tukey multiple comparisons of means

# 99% family-wise confidence level

#

# Fit: aov(formula = center\_table$public.health.hazard.violation.rate ~ center\_table$borough)

#

# $`center\_table$borough`

# diff lwr upr p adj

# BROOKLYN-BRONX -11.1786965 -15.520631 -6.83676230 0.0000000

# MANHATTAN-BRONX -5.8410530 -10.443683 -1.23842330 0.0003529

# QUEENS-BRONX -4.9648821 -9.665530 -0.26423386 0.0053200

# STATEN ISLAND-BRONX -7.2412279 -14.579728 0.09727231 0.0115424

# MANHATTAN-BROOKLYN 5.3376435 1.689659 8.98562807 0.0000194

# QUEENS-BROOKLYN 6.2138144 2.442915 9.98471418 0.0000009

# STATEN ISLAND-BROOKLYN 3.9374686 -2.843179 10.71811582 0.3219232

# QUEENS-MANHATTAN 0.8761709 -3.192182 4.94452352 0.9561708

# STATEN ISLAND-MANHATTAN -1.4001749 -8.350641 5.55029161 0.9654999

# STATEN ISLAND-QUEENS -2.2763458 -9.292105 4.73941346 0.8284439

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: tbl2

# X-squared = 267.08, df = 104, p-value = 2.665e-16

#######As the p-value is less than the .05 significance level, we reject the null hypothesis that ######the PHH violation rate is independent of the borough.

1. Borough and Critical Violation Rate
2. One-Way ANOVA Test

#Df Sum Sq Mean Sq F value Pr(>F)

#center\_table$borough 4 103806 25951 35.66 <2e-16 \*\*\*

# Residuals 2860 2081105 728

#---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Tukey’s Honestly Significant Differences

# Tukey multiple comparisons of means

# 99% family-wise confidence level

#

# Fit: aov(formula = center\_table$critical.violation.rate ~ center\_table$borough)

#

# $`center\_table$borough`

# diff lwr upr p adj

# BROOKLYN-BRONX -16.795517 -21.956488 -11.634546 0.0000000

# MANHATTAN-BRONX -18.091978 -23.562821 -12.621135 0.0000000

# QUEENS-BRONX -14.495792 -20.083143 -8.908441 0.0000000

# STATEN ISLAND-BRONX -19.673418 -28.396210 -10.950627 0.0000000

# MANHATTAN-BROOKLYN -1.296461 -5.632580 3.039658 0.8669586

# QUEENS-BROOKLYN 2.299725 -2.182495 6.781945 0.4518701

# STATEN ISLAND-BROOKLYN -2.877902 -10.937610 5.181807 0.7723550

# QUEENS-MANHATTAN 3.596186 -1.239596 8.431969 0.1095294

# STATEN ISLAND-MANHATTAN -1.581440 -9.843002 6.680121 0.9713358

# STATEN ISLAND-QUEENS -5.177627 -13.516798 3.161544 0.2552626

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: tbl3

# X-squared = 403.94, df = 124, p-value < 2.2e-16

#######As the p-value is less than the .05 significance level, we reject the null hypothesis that ######the critical violation rate is independent of the borough.

**Appendix B.**

1. Facility Type and Violation Rate Percent
2. One-way ANOVA Test

# Df Sum Sq Mean Sq F value Pr(>F)

# center\_table$facility.type 2 11487 5744 6.901 0.00102 \*\*

# Residuals 2862 2382016 832

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#####The F-statistic value is 6.901and it is highly signficiant as the p-value is

####much less than the level of significance (1% or 0.01). Therefore, reject the

####null hypothesis of equal mean value of violation rate percent across all the

####facility types.

1. Tukey’s Honestly Significant Differences

# 99% family-wise confidence level

#

# Fit: aov(formula = center\_table$violation.rate.percent ~ center\_table$facility.type)

#

# $`center\_table$facility.type`

# diff lwr upr p adj

# GDC-CAMP 1.309662 -2.836137 5.455461 0.6269646

# SBCC-CAMP 6.513450 1.004182 12.022717 0.0016638

# SBCC-GDC 5.203788 0.672616 9.734959 0.0023707

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: facil\_tbl

# X-squared = 938.15, df = 64, p-value < 2.2e-16

#

# Warning message:

# In chisq.test(facil\_tbl) : Chi-squared approximation may be incorrect

#######As the p-value is less than the .05 significance level, we reject the null hypothesis that ######the violation rate percent is independent of the facility types.

1. Program Type and Violation Rate Percent
2. One-Way ANOVA Test

# Df Sum Sq Mean Sq F value Pr(>F)

# center\_table$program.type 3 3855 1285.1 1.539 0.202

# Residuals 2861 2389649 835.2

#####The F-statistic value is 1.539 and the p-value is 0.202. Therefore, do not reject the

####null hypothesis of equal mean value of violation rate percent across all the

####program types.

1. Tukey’s Honestly Significant Differences

# Tukey multiple comparisons of means

# 99% family-wise confidence level

#

# Fit: aov(formula = center\_table$violation.rate.percent ~ center\_table$program.type)

#

# $`center\_table$program.type`

# diff lwr upr p adj

# INFANT TODDLER-ALL AGE CAMP 0.3778064 -5.780774 6.536386 0.9975247

# PRESCHOOL-ALL AGE CAMP 2.3081965 -2.224560 6.840953 0.3863298

# SCHOOL AGE CAMP-ALL AGE CAMP -4.2959971 -21.229516 12.637521 0.8587930

# PRESCHOOL-INFANT TODDLER 1.9303901 -3.133275 6.994055 0.6345652

# SCHOOL AGE CAMP-INFANT TODDLER -4.6738034 -21.757096 12.409489 0.8291922

# SCHOOL AGE CAMP-PRESCHOOL -6.6041936 -23.170927 9.962539 0.5999884

#Tukey’s HSD shows that all the pairs of violation rate percent are statistically significantly different except for the pari Infant Toddler-All Age camp

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: prog\_tbl

# X-squared = 575.18, df = 96, p-value < 2.2e-16

## As the p-value is less than .05 significance level, we reject the null hypothesis that the violation rate percent is independent of program types.

**Appendix C.**

1. Maximum Capacity and Violation Rate Percent
2. Correlation between maximum capacity and violation rate percent

# Pearson's product-moment correlation

#

# data: viol\_rate and center\_table$maximum.capacity

# t = -1.9618, df = 2863, p-value = 0.04988

# alternative hypothesis: true correlation is not equal to 0

# 95 percent confidence interval:

# -7.316130e-02 -1.937356e-05

# sample estimates:

# cor

# -0.03663941

1. One-way ANOVA Test

# Df Sum Sq Mean Sq F value Pr(>F)

# center\_table$borough 4 34401 8600 0.514 0.725

# Residuals 2860 47855377 16733

#####The F-statistic value is 0.514 and the p-value is 0.725. Therefore, do not reject the

####null hypothesis of equal mean value of maximum capacity across all the

####boroughs.

1. Tukey’s Honestly Significant Differences

# Tukey multiple comparisons of means

# 99% family-wise confidence level

#

# Fit: aov(formula = center\_table$maximum.capacity ~ center\_table$borough)

#

# $`center\_table$borough`

# diff lwr upr p adj

# BROOKLYN-BRONX 4.4131907 -20.33536 29.16174 0.9779025

# MANHATTAN-BRONX -4.3963609 -30.63085 21.83813 0.9824519

# QUEENS-BRONX 0.4571268 -26.33605 27.25031 0.9999978

# STATEN ISLAND-BRONX 4.8262429 -37.00241 46.65489 0.9957617

# MANHATTAN-BROOKLYN -8.8095515 -29.60266 11.98356 0.6404500

# QUEENS-BROOKLYN -3.9560639 -25.44978 17.53765 0.9751700

# STATEN ISLAND-BROOKLYN 0.4130522 -38.23590 39.06200 0.9999997

# QUEENS-MANHATTAN 4.8534876 -18.33568 28.04265 0.9604171

# STATEN ISLAND-MANHATTAN 9.2226037 -30.39430 48.83950 0.9423486

# STATEN ISLAND-QUEENS 4.3691161 -35.61994 44.35818 0.9965693

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: maxcap\_tbl

# X-squared = 8468.8, df = 8960, p-value = 0.9999

1. Total Educational Workers and Violation Rate Percent
2. Summary

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 0.000 1.000 6.000 8.353 12.000 115.000

1. Correlation test

# Pearson's product-moment correlation

#

# data: viol\_rate and center\_table$total.educational.workers

# t = 1.8497, df = 2863, p-value = 0.06446

# alternative hypothesis: true correlation is not equal to 0

# 95 percent confidence interval:

# -0.002074591 0.071078225

# sample estimates:

# cor

# 0.03454809

1. Chi-squared Test

# Pearson's Chi-squared test

#

# data: staff\_tbl

# X-squared = 2925.1, df = 1952, p-value < 2.2e-16

#

# Warning message:

# In chisq.test(staff\_tbl) : Chi-squared approximation may be incorrect

#######As the p-value is less than the .05 significance level,

# we reject the null hypothesis that the violation rate percent is independent

# of the number of educational workers. In other words, we favor the null hypothesis that

#the violation rate percent is affected by the number of educational workers.