Will New York City Kill Your Children

Roshni Patnayakuni, Olivia Handoko, Truc Tran

12/20/2019

Abstract:

This research focused on the variables that can potentially affect whether a childcare center will have a health violation. 2196 childcare centers were observed in this analysis across a span of 3 years. We predicted that longer standing childcares may have lower critical violation rate while also hypothesizing that zip code might be correlated with the amount of critical violation rate as well. Therefore, there could be higher critical violations when combining zip code and date permitted as interaction terms for those in low socioeconomic neighborhoods and child cares that may have not lasted as long as others. As well, we hypothesize that since zip code may have an effect on the conditions of a childcare center, educational workers may be correlated to the critical violation rate, as the more workers there are at a childcare center, as this may be indicative of the economic status of the business. Combining all 3 possible critical predictors would create the most variability in the effects it has on critical violation rates. Our analysis showed that with all 5 predictors ("Critical Violation Rate", "Maximum Capacity", "Total Education Workers", "Date Permitted", " Zip Code", and "Program Type"), there showed little variance in explaining the response variable, critical violation rate (<1%).

Introduction

New York City hosts approximately 8.623 million people (from the 2017 consensus) which makes the city the most populous place in the United States. Of those statistics, more than 1.7 million children make up New York City's population. That's about 20% of the population! It is no doubt that with that many residents, New York City may struggle with cleanliness. However, these health inspections are mostly talked about in food facilities and rarely are these about facilities that cater to children. Childcare centers across New York City are often one of the main places that specifically serves children and an interesting target to research the cleanliness of New York City. For our research project, we decided to analyze New York City's Department of Health and Mental Hygiene. This dataset was offered through Kaggle's free database that allowed us to see important statistics regarding the conditions of all the childcare centers in file, from December 2016 to December 2019, after inspections. Within this dataset, there are a total of around 53,400 observations with 2,196 unique child care centers all across New York City. Of those columns identified, we decided to focus on "Critical Violation Rate", "Maximum Capacity", "Total Education Workers", "Date Permitted", " Zip Code", and "Program Type" as critical predictors in influencing lower critical violation rate in these centers. With a focus on those variables, we wanted to know which of the 5 predictors has the most significant effect on lower critical violation rate as our main research question. To do this, we created a multiple regression model that would give us a sense of the 5 predictors in explaining the results of the response variable (critical violation rate). Our results from this analysis will help us to predict where violations may happen to better prevent childcare centers from having them in the future.

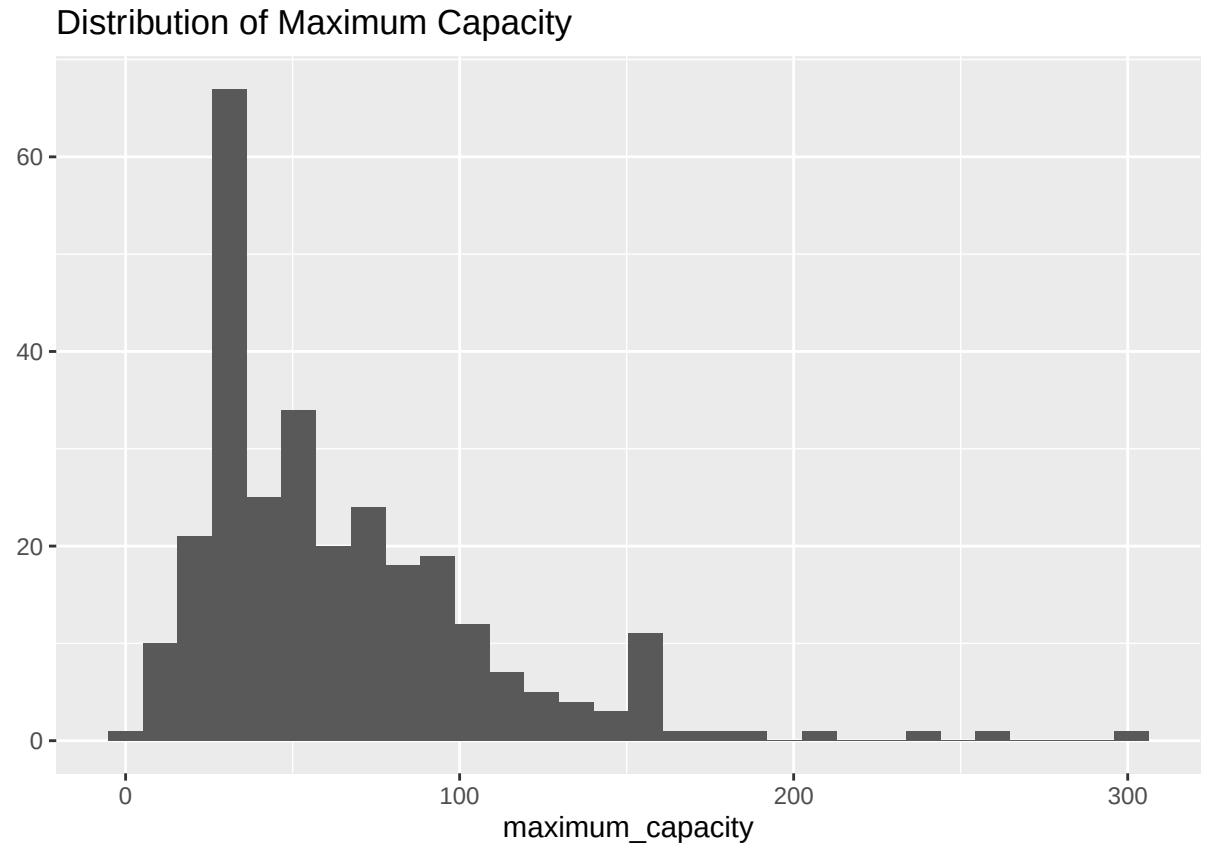
Data

Our dataset gives information on Childcare Center Inspections in New York City and comes from the Department of Health and Mental Hygiene. The dataset contains the information gathered from inspections dating back to December 2016 all the way up to December 2019 and contains a variety of information about the centers themselves and the types and frequency of violations. It includes data from approximately 53,400 inspections across 2196 childcare centers. For our research project we examined the Critical Violation Rate, the percent of total critical violations broken by the childcare center at each inspection, as our response.

Critical Violations are second degree violations: less severe then Public Health Hazards but more serious than General Violations., and must be corrected within 2 weeks. For our explanatory variables we primarily focused on data about the childcare centers themselves, so we can use the results to understand what types of centers might be considered unsafe. Specifically we examined the maximum capacity: the largest number of children the center can host, the total educational workers: the number of staff members involved in education, The date permitted, which will be used to determine the age of each day care in days, the zip code of the day care and the program type: whether they are an infant/toddler daycare or a preschool.

```
qplot(maximum_capacity, data = data, main = "Distribution of Maximum Capacity")
```

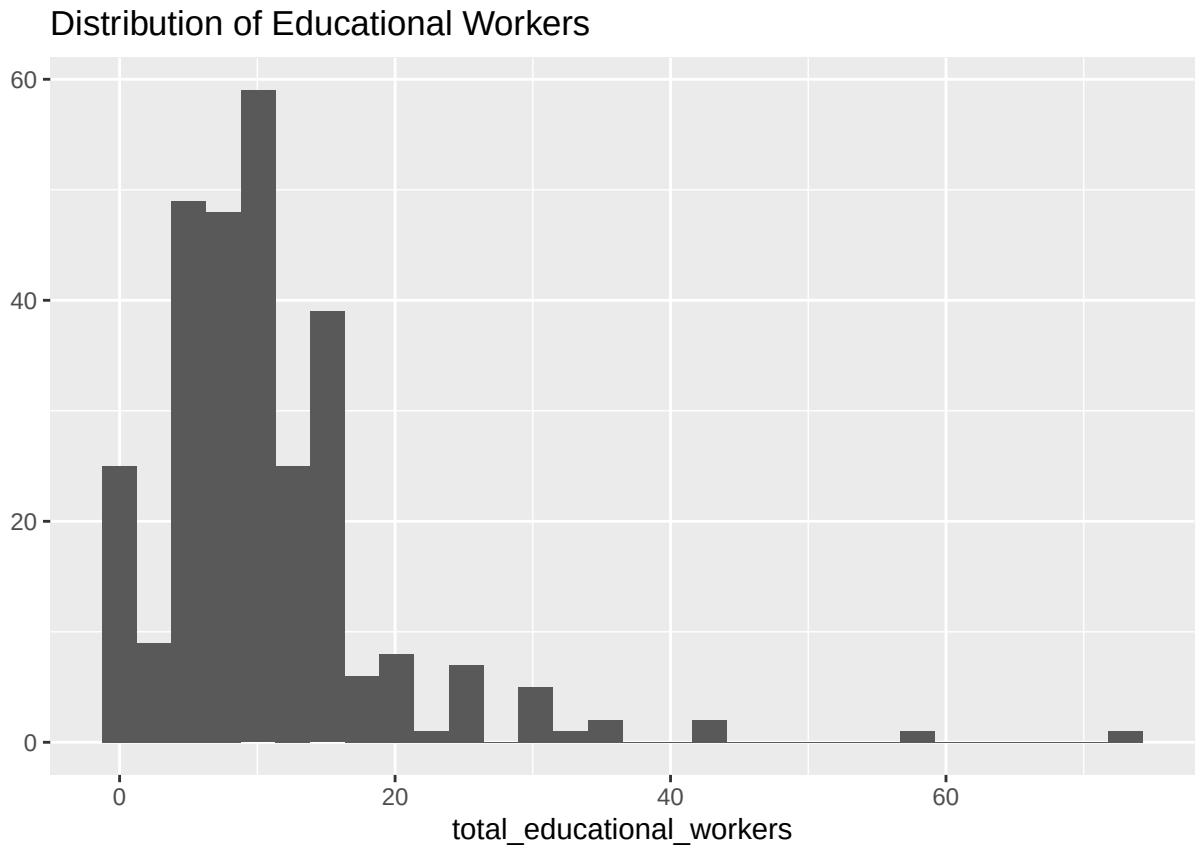
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram of maximum capacity indicates that the distribution of this variable skews slightly to the left, as well as indicating some points with unusually high maximum capacity.

```
qplot(total_educational_workers, data = data, main = "Distribution of Educational Workers")
```

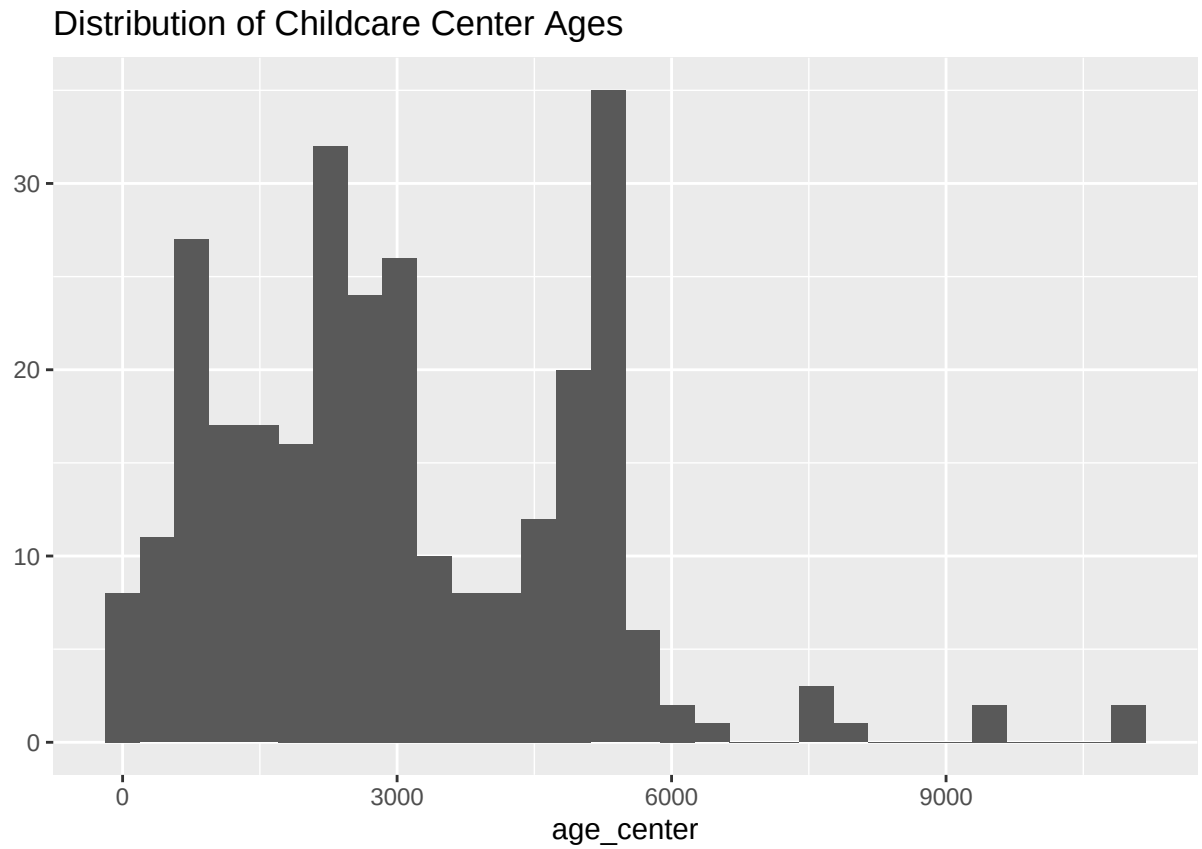
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Again, the histogram of total educational workers indicates that the distribution of this variable skews slightly to the left, as well as indicating some points with unusually high number of educational workers.

```
qplot(age_center, data = data, main = "Distribution of Childcare Center Ages")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Unlike the previous two variables, age of childcare center does not have a distribution that can be labelled as right or left skewed. Instead there appear to be 3 peaks centerd at ages of approximately 750 days, 2250 days, and 5250 days.

Results

```
regModel <- lm(critical_violation_rate ~ zip_code + program_type + maximum_capacity + total_educational_

summary(regModel)
```

```
##
## Call:
## lm(formula = critical_violation_rate ~ zip_code + program_type +
##     maximum_capacity + total_educational_workers + age_center,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.472 -17.215   0.257  16.037  65.033
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               29.6383299 27.2413756   1.088   0.2775
## zip_code                   0.0011437  0.0025481   0.449   0.6539
## program_typePRESCHOOL     -6.1334404  4.0801253  -1.503   0.1339
## maximum_capacity           0.0724998  0.0454887   1.594   0.1121
```

```
## total_educational_workers -0.1603867  0.2255495  -0.711   0.4776
## age_center                -0.0011862  0.0006677  -1.776   0.0767 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.81 on 282 degrees of freedom
## Multiple R-squared:  0.02494,    Adjusted R-squared:  0.007653
## F-statistic: 1.443 on 5 and 282 DF,  p-value: 0.209
```

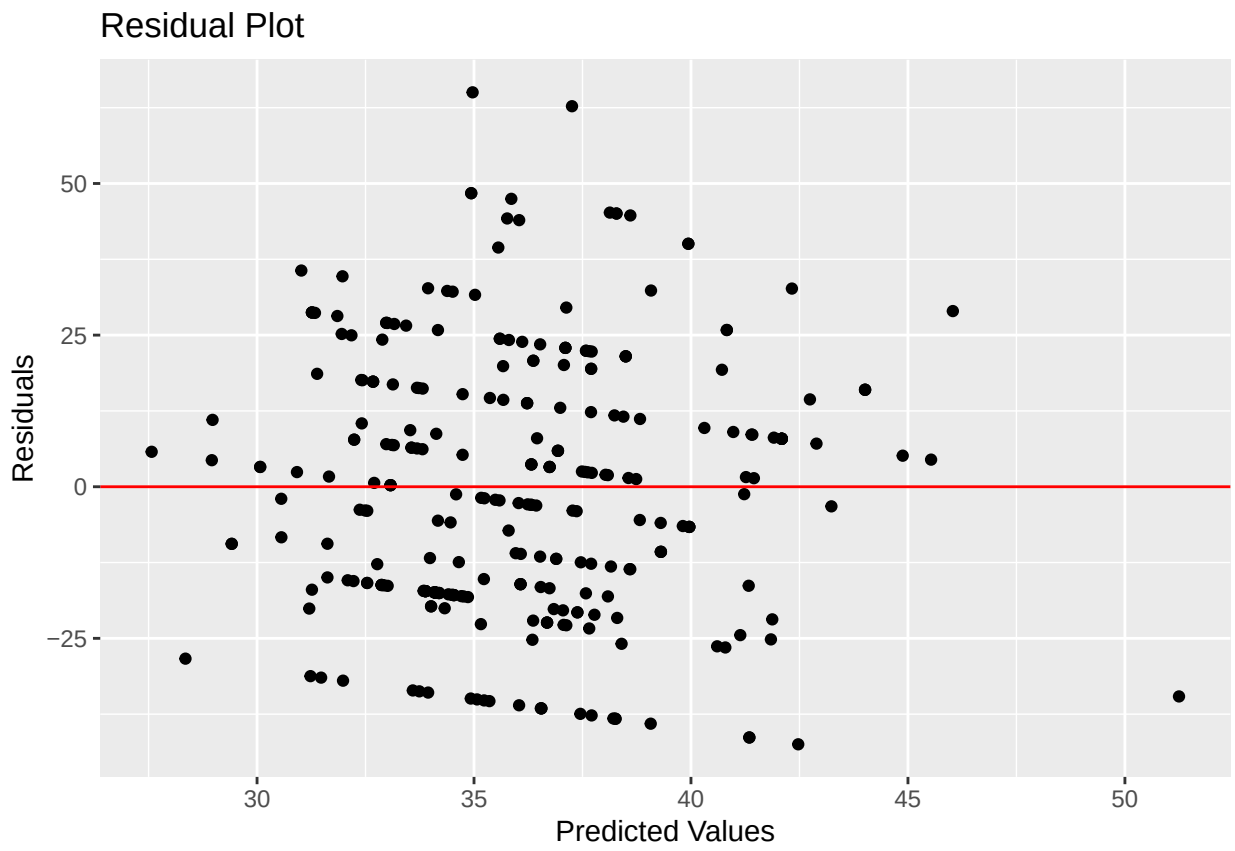
The results from the model summary indicate that none of the variable cause significant change to critical violation rate, as they all have p values greater than 0.05. Furthermore, the adjusted r-squared indicates that this model only accounts for 0.77% of the variance in critical violation rate.

```
anova(regModel)
```

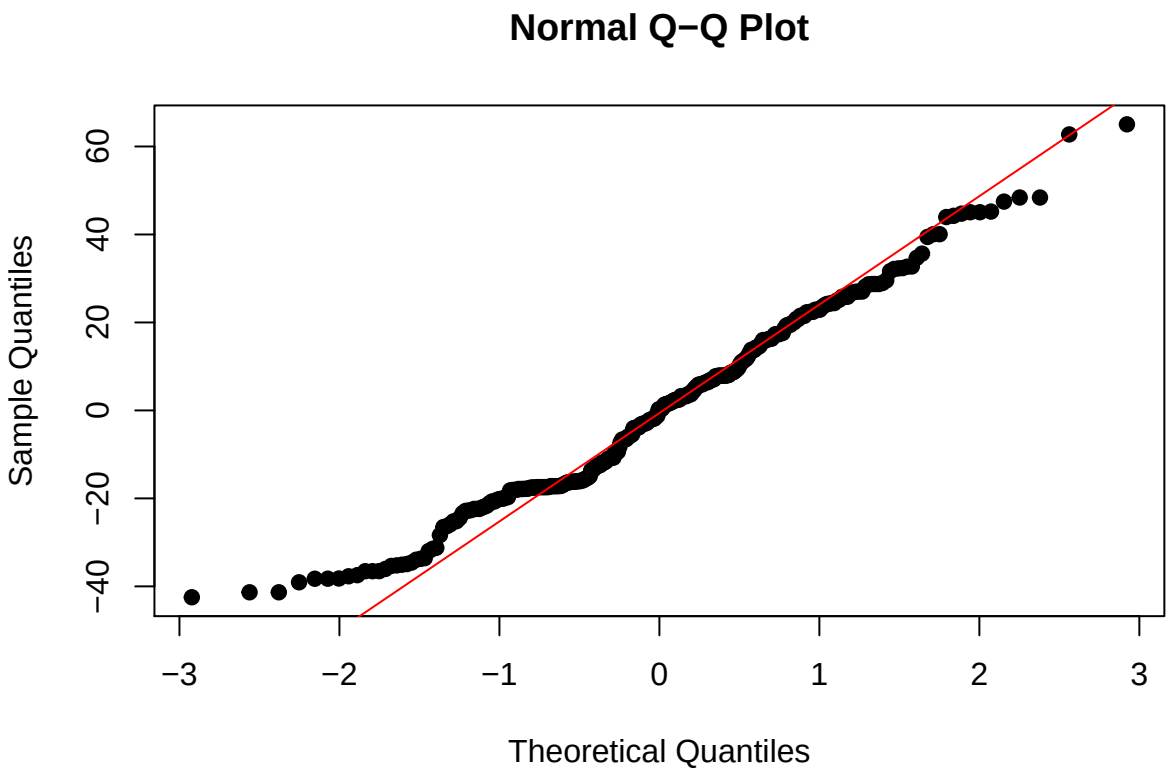
```
## Analysis of Variance Table
##
## Response: critical_violation_rate
##                            Df Sum Sq Mean Sq F value  Pr(>F)
## zip_code                    1      0    0.24  0.0005 0.98195
## program_type                1    669  668.59  1.4049 0.23690
## maximum_capacity            1    976  976.02  2.0509 0.15322
## total_educational_workers   1    286  286.09  0.6012 0.43878
## age_center                  1   1502 1501.89  3.1559 0.07673 .
## Residuals                 282 134202  475.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance table reinforces what was learned from the coefficients table. None of the coefficents have a significant effect on the response variable.

```
predVal <- predict(regModel)
residVal <- residuals(regModel)
ggplot(mapping = aes(x = predVal, y = residVal)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residual Plot", x = "Predicted Values", y = "Residuals")
```




```
qqnorm(residVal, pch=19)
qqline(residVal, col="red")
```



Upon Assessing the conditions of linearity it seems that most but not all the conditions of linearity are fulfulled. The residual plot indicates that while the model is linear and has constant variance the residuals are not independent to the predicted Y value. The Q-Q plot indicates that the condition of normality is fufilled for the model.

Most Significant Variable

```
m_zip <- lm(critical_violation_rate ~ zip_code, data = data)
m_prog <- lm(critical_violation_rate ~ program_type, data = data)
m_max <- lm(critical_violation_rate ~ maximum_capacity, data = data)
m_edWorker <- lm(critical_violation_rate ~ total_educational_workers, data = data)
m_age <- lm(critical_violation_rate ~ age_center, data = data)

summary(m_zip)
```

```
##
## Call:
## lm(formula = critical_violation_rate ~ zip_code, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.121 -19.402  -2.711  13.956  63.891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.548e+01  2.676e+01   1.326    0.186
```

```
## zip_code    5.580e-05  2.478e-03   0.023    0.982
##
## Residual standard error: 21.94 on 286 degrees of freedom
## Multiple R-squared:  1.774e-06,  Adjusted R-squared:  -0.003495
## F-statistic: 0.0005073 on 1 and 286 DF,  p-value: 0.982
```

```
summary(m_prog)
```

```
##
## Call:
## lm(formula = critical_violation_rate ~ program_type, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.814 -18.818  -2.151  14.515  64.515
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             39.814      3.460  11.506   <2e-16 ***
## program_typePRESCHOOL   -4.330      3.729  -1.161    0.247
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.89 on 286 degrees of freedom
## Multiple R-squared:  0.004692,   Adjusted R-squared:  0.001212
## F-statistic: 1.348 on 1 and 286 DF,  p-value: 0.2466
```

```
summary(m_max)
```

```
##
## Call:
## lm(formula = critical_violation_rate ~ maximum_capacity, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.873 -18.420  -1.972  14.957  65.030
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      34.21286    2.28274  14.988   <2e-16 ***
## maximum_capacity  0.02913    0.02928   0.995    0.321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.9 on 286 degrees of freedom
## Multiple R-squared:  0.003448,   Adjusted R-squared:  -3.629e-05
## F-statistic: 0.9896 on 1 and 286 DF,  p-value: 0.3207
```

```
summary(m_edWorker)
```

```
##
## Call:
## lm(formula = critical_violation_rate ~ total_educational_workers,
##     data = data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -38.328 -18.721  -2.489  14.228  64.345
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               35.38817    2.04979  17.264   <2e-16 ***
## total_educational_workers  0.06682    0.15234   0.439    0.661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.93 on 286 degrees of freedom
## Multiple R-squared:  0.0006721,  Adjusted R-squared:  -0.002822
## F-statistic: 0.1924 on 1 and 286 DF,  p-value: 0.6613
```

```
summary(m_age)
```

```
##
## Call:
## lm(formula = critical_violation_rate ~ age_center, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.660 -19.197  -0.296  16.527  66.599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.711204   2.375480  16.717   <2e-16 ***
## age_center  -0.001196   0.000659  -1.815   0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.81 on 286 degrees of freedom
## Multiple R-squared:  0.01138,    Adjusted R-squared:  0.007926
## F-statistic: 3.293 on 1 and 286 DF,  p-value: 0.07063
```

Upon examining the effect of each explanatory variable individually it can be determined that age_center has the greatest effect on critical violation rate. This aligns with the previous results that indicate that of all the variables used, the age of the center has the highest significance.

Finding Unusual Points

```
# Diagnostic
diag <- ls.diag(regModel)

unusual_points <- data %>%
  mutate(h_i = diag$hat,
         stnd_res = diag$std.res,
         stud_res = diag$stud.res,
         cooks = diag$cooks)
# H_i
unusual_points %>%
  filter(h_i > 12/43538  | h_i > 18/43538) %>%
  head(5)
```

```
##   zip_code maximum_capacity program_type total_educational_workers
```

```
## 1    10468               30    PRESCHOOL                        0
## 2    10468               30    PRESCHOOL                        0
## 3    10468               30    PRESCHOOL                        0
## 4    10468               30    PRESCHOOL                        0
## 5    10468               30    PRESCHOOL                        0
##   critical_violation_rate date_permitted inspection_date age_center
## 1                 16.6667     2010-12-06      2018-02-14  2627 days
## 2                 16.6667     2010-12-06      2017-08-07  2436 days
## 3                 16.6667     2010-12-06      2018-02-14  2627 days
## 4                 16.6667     2010-12-06      2018-02-14  2627 days
## 5                 16.6667     2010-12-06      2017-09-22  2482 days
##          h_i   stnd_res   stud_res       cooks
## 1 0.01129142 -0.8238159 -0.8233452 0.001291782
## 2 0.01144249 -0.8343240 -0.8338732 0.001342879
## 3 0.01129142 -0.8238159 -0.8233452 0.001291782
## 4 0.01129142 -0.8238159 -0.8233452 0.001291782
## 5 0.01139986 -0.8317905 -0.8313348 0.001329705
```

```
# Standardized residual
unusual_points %>%
  filter(abs(stnd_res) > 2 | abs(stnd_res) > 3) %>%
  head(5)
```

```
##   zip_code maximum_capacity program_type total_educational_workers
## 1    10466              120    PRESCHOOL                        21
## 2    10466              120    PRESCHOOL                        21
## 3    10466              120    PRESCHOOL                        21
## 4    11203               56    PRESCHOOL                         8
## 5    11436               95    PRESCHOOL                        14
##   critical_violation_rate date_permitted inspection_date age_center
## 1                83.3333     2012-04-13      2018-02-09  2128 days
## 2                83.3333     2012-04-13      2018-02-09  2128 days
## 3                83.3333     2012-04-13      2017-05-15  1858 days
## 4                83.3333     2015-07-28      2017-10-18   813 days
## 5               100.0000     2003-11-12      2018-04-24  5277 days
##          h_i stnd_res stud_res       cooks
## 1 0.01285891 2.078521 2.090911 0.009379565
## 2 0.01285891 2.078521 2.090911 0.009379565
## 3 0.01357142 2.064490 2.076579 0.009773138
## 4 0.01093822 2.083519 2.096017 0.008001413
## 5 0.01538317 3.004300 3.048146 0.023502501
```

```
# Studentized residual
unusual_points %>%
  filter(abs(stud_res) > 2 | abs(stud_res) > 3) %>%
  head(5)
```

```
##   zip_code maximum_capacity program_type total_educational_workers
## 1    10466              120    PRESCHOOL                        21
## 2    10466              120    PRESCHOOL                        21
## 3    10466              120    PRESCHOOL                        21
## 4    11203               56    PRESCHOOL                         8
## 5    11436               95    PRESCHOOL                        14
##   critical_violation_rate date_permitted inspection_date age_center
## 1                83.3333     2012-04-13      2018-02-09  2128 days
## 2                83.3333     2012-04-13      2018-02-09  2128 days
```

```
## 3                 83.3333     2012-04-13      2017-05-15  1858 days
## 4                 83.3333     2015-07-28      2017-10-18   813 days
## 5                100.0000     2003-11-12      2018-04-24  5277 days
##           h_i stnd_res stud_res       cooks
## 1 0.01285891 2.078521 2.090911 0.009379565
## 2 0.01285891 2.078521 2.090911 0.009379565
## 3 0.01357142 2.064490 2.076579 0.009773138
## 4 0.01093822 2.083519 2.096017 0.008001413
## 5 0.01538317 3.004300 3.048146 0.023502501
```

These results indicat that there are a large number of points that could be considered outliers as follows: Leverage: 2876 unusual points Standardized residual: 1038 unusual points Studentized residual; 1038 unusual points.

Conclusion

This model does not do a good job of predicting the change in critical violation rate, as it only explains a small portion of variance in the response variable (<1%) and none of the variables are statistically significant. As well the model did not fulfill all the conditions of regression. If this project were to be continued, it might be worthwile to reproduce the model without the outliers listed above. As well, it may be helpful to test a transformation on some of the variables to fix the independence of residuals. Ultimately if none of this worked, it would be beneficial to look into other variables, perhaps those that present information on the results of the inspection, instead of the profile of the center.