PA446 Final Project Report

Haedodam Kim

2025-12-10

Table of contents

1 1. Introduction

Arts education plays a critical role in supporting students' academic achievement, social, emotional development, and long-term civic engagement. Despite its importance, access to high-quality arts programming in Chicago Public Schools (CPS) varies substantially across neighborhoods, raising persistent concerns about educational equity.

To evaluate arts education environments and program quality, CPS operates the Creative Schools Certification (CSC) system, which categorizes schools into four levels: Emerging, Developing, Strong, and Excelling. While CSC provides a useful benchmark for school-level arts access, there has been limited analysis of how these scores relate to broader socioeconomic conditions across ZIP codes, and which communities may be structurally disadvantaged in receiving robust arts education.

This project examines **how neighborhood-level socioeconomic factors, such as poverty rate, racial composition, linguistic diversity, and educational attainment, are associated with CSC outcomes across Chicago ZIP codes**. Using spatial visualization, the analysis identifies geographic patterns of inequality, and applies a decision tree classifier to explore which factors most strongly differentiate high- and low-scoring areas.

The goal of this study is to provide a data-driven understanding of inequities in arts education across Chicago, and to offer empirical insights for identifying which neighborhoods require targeted support, as well as potential policy strategies to advance arts education equity city-wide.

2 2. Data Acquisition

This analysis draws on two primary data sources to examine the relationship between neighborhood socioeconomic conditions and arts education access across Chicago.

First, I obtained ZIP Code-level indicators from the **American Community Survey (ACS) 2022 5-year estimates** using the API. Key variables included:

- Poverty rate
- Percent White (used to derive percent non-White)
- Percent English-only households (used to derive percent non-English-speaking)
- Percent of residents with a bachelor's degree or higher (used to derive percent non-college)

Second, I used the **Creative Schools Certification (CSC)** dataset provided through **Ingenuity's ArtLook Maps**, a platform that documents CPS arts education resources, programs, and partnerships. ZIP codes were extracted from school address fields, and CSC levels (Emerging, Developing, Strong, Excelling) were converted into a numeric scale ranging from 1 to 4 to enable quantitative analysis.

Finally, the ACS and ArtLook datasets were merged by ZIP code to create a unified dataset linking neighborhood socioeconomic conditions with arts education outcomes across Chicago.

3 3. Data Wrangling & Quality Checks

After merging the datasets, I conducted basic cleaning and quality checks to ensure the reliability of the analysis. I first reviewed variable structures and summary statistics to identify any data type issues or unusual values. Missingness was then examined using `gg_miss_var()` and `vis_miss()` to check whether CSC scores or key socioeconomic indicators had missing values or showed any systematic patterns.

I also counted the number of schools within each ZIP code to assess representativeness, since ZIPs with very few observations may produce unstable averages. Finally, rows missing essential variables (CSC score, poverty rate, and the derived racial, language, and education indicators) were removed to create a consistent dataset for analysis.

4 4. Analysis

This section investigates three core analytical questions:
1. Are arts education levels (CSC scores) evenly distributed across Chicago?
2. Which socioeconomic factors are associated with CSC scores?
3. How do CSC scores relate to each factor individually?

```
# Load required packages
library(tidyverse)
library(tidycensus)
library(rvest)
library(stringr)
library(naniar)
library(broom)
library(tigris)
library(sf)
library(ggplot2)
library(shiny)
library(DT)
library(bslib)
library(rpart)
library(rpart.plot)

options(tigris_use_cache = TRUE)
```

```
# Load cleaned datasets
arts_full <- readr::read_csv("../data/arts_full.csv")
arts_full <- arts_full %>% mutate(zip = as.character(zip))
```

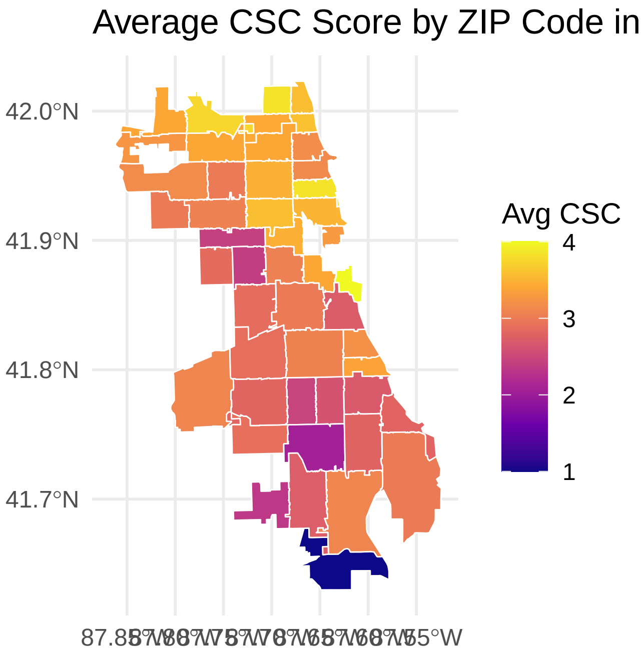
1. Are arts education levels (CSC scores) evenly distributed across Chicago?

```
# Aggregate to ZIP level
zip_summary <- arts_full %>%
  group_by(zip) %>%
  summarize(
    avg_CSC         = mean(CSC_score, na.rm = TRUE),
    poverty_rate    = unique(poverty_rate),
    pct_nonwhite   = unique(pct_nonwhite),
    pct_non_english= unique(pct_non_english),
    pct_noncollege = unique(pct_noncollege),
    n_schools      = n()
  )

# Load Chicago ZCTA shapefile (2020)
chi_zcta <- zctas(cb = TRUE, year = 2020) %>%
  mutate(zip = as.character(ZCTA5CE20)) %>%
  filter(zip %in% zip_summary$zip)

# Merge spatial data with aggregated ZIP summary
chi_map <- chi_zcta %>%
  left_join(zip_summary, by = "zip")

# Map: Visualize spatial distribution of CSC scores
ggplot(chi_map) +
  geom_sf(aes(fill = avg_CSC), color = "white", size = 0.25) +
  scale_fill_viridis_c(option = "plasma") +
  labs(
    title = "Average CSC Score by ZIP Code in Chicago",
    fill  = "Avg CSC"
  ) +
  theme_minimal()
```


Average CSC Score by ZIP Code in Chicago
42.0°N
41.9°N
41.8°N
41.7°N
87.85°W87.80°W87.75°W87.70°W87.65°W87.60°W87.55°W
Avg CSC
4
3
2
1

CSC scores show clear spatial inequality across Chicago: higher-scoring ZIP codes are concentrated in the North Side and near Downtown, while lower-scoring areas cluster in the South and West Sides.

2. Which socioeconomic factors are associated with CSC scores?

```
# Regression Model
model_all <- lm(
  avg_CSC ~ poverty_rate + pct_nonwhite + pct_non_english + pct_noncollege,
  data = zip_summary
)

# Print regression output
summary(model_all)
```

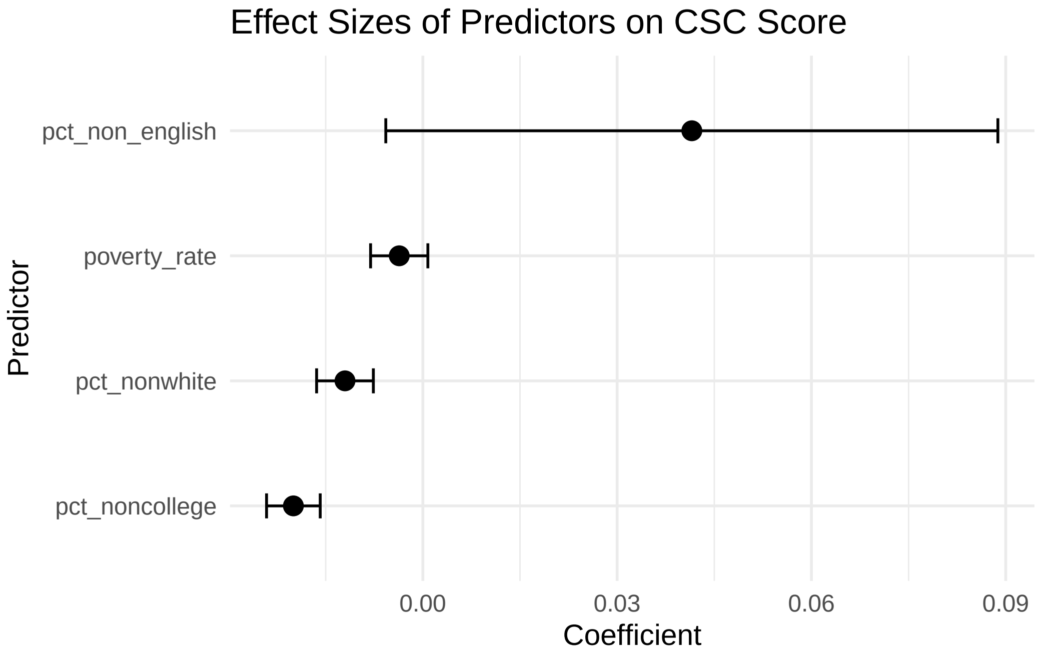
```
Call:
lm(formula = avg_CSC ~ poverty_rate + pct_nonwhite + pct_non_english +
    pct_noncollege, data = zip_summary)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42386 -0.13658  0.03302  0.20623  0.60708
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.308776   4.587828   0.285  0.77681    
poverty_rate    -0.003646   0.004422  -0.825  0.41418    
pct_nonwhite    -0.012021   0.004382  -2.743  0.00883 ** 
pct_non_english  0.041532   0.047254   0.879  0.38433    
pct_noncollege  -0.019988   0.004141  -4.827 1.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3642 on 43 degrees of freedom
Multiple R-squared:  0.5491,    Adjusted R-squared:  0.5072 
F-statistic: 13.09 on 4 and 43 DF,  p-value: 4.669e-07
```

```r
# Visualize Regression Coefficients
tidy(model_all) %>%
  filter(term != "(Intercept)") %>%
  ggplot(aes(x = reorder(term, estimate), y = estimate)) +
  geom_point(size = 3) +
  geom_errorbar(
    aes(ymin = estimate - std.error,
        ymax = estimate + std.error),
    width = 0.2
  ) +
  coord_flip() +
  labs(
    title = "Effect Sizes of Predictors on CSC Score",
    x = "Predictor",
    y = "Coefficient"
  ) +
  theme_minimal()
```



This regression coefficient plot illustrates the associations between ZIP-level CSC scores and key socioeconomic factors. The results show that the percentage of non-English-speaking residents is the only variable with a clear positive effect, indicating the strongest relationship with CSC scores. In contrast, poverty rate, racial composition, and educational attainment display small effect sizes with confidence intervals that cross zero, suggesting that these factors are not strong predictors of arts education levels.

3. How do CSC scores relate to each factor individually?

```r
# Interactive Shiny Dashboard
# The dashboard allows:
# - selecting a socioeconomic predictor for scatter plot
# - viewing first N ZIP rows in a data table
# - personalized greeting using user name input

df <- zip_summary   # ZIP-level dataset for dashboard

# Numeric predictors available in scatter plot
num_cols <- c("poverty_rate", "pct_nonwhite", "pct_non_english", "pct_noncollege")

# UI layout
ui <- fluidPage(
  theme = bs_theme(version = 5, bootswatch = "flatly"),
  div(style = "padding: 12px 0; font-weight: 600; font-size: 1.2rem;",
      textOutput("greet")
  ),

  sidebarLayout(
    sidebarPanel(
      h4("Controls"),
      textInput("name", "What is your name?"),

      selectInput(
        "xvar",
        "Select predictor (X-axis):",
        choices = num_cols,
        selected = "pct_nonwhite"
      ),

      numericInput(
        "nrows",
        "Number of ZIP rows to show:",
        value = 10, min = 1, max = nrow(df)
      )
    ),

    mainPanel(
      tabsetPanel(
        tabPanel("Scatter Plot", plotOutput("scatter", height = 420)),
        tabPanel("ZIP Table", DTOutput("table"))
      )
    )
  )
)

# SERVER logic
server <- function(input, output, session) {

  # Greeting message updates depending on user input
  output$greet <- renderText({
    nm <- trimws(input$name)
    if (nm == "" || is.null(nm))
      "Welcome to the Chicago Arts Edu Equity Explorer!"
    else
      paste0("Hello, ", nm, "! Explore how local conditions shape CSC scores.")
  })

  # Scatter plot (Y-axis fixed as avg_CSC)
  output$scatter <- renderPlot({
    ggplot(df, aes(x = .data[[input$xvar]], y = avg_CSC)) +
      geom_point(size = 3, alpha = 0.8, color = "#0072B2") +
      geom_smooth(method = "lm", color = "red", se = TRUE) +
      labs(
        x = input$xvar,
        y = "Average CSC Score",
        title = "Relationship Between CSC Score and Socioeconomic Factors"
      ) +
      theme_minimal(base_size = 14)
  })

  # Display first N rows of ZIP summary data
  output$table <- renderDT({
    datatable(
      df %>% head(input$nrows),
      rownames = FALSE,
      options = list(pageLength = input$nrows)
    )
  })
}

# RUN THE APP
shinyApp(ui, server)
```

The dashboard shows that CSC scores do not exhibit strong linear relationships with any single socioeconomic factor. While there are minor regional differences and generally weak associations overall, none of the ZIP-level indicators serve as strong standalone predictors of arts education quality in Chicago. However, the share of residents without a college degree shows a slight positive trend, suggesting a potential but modest association that may relate to local educational or cultural access patterns.

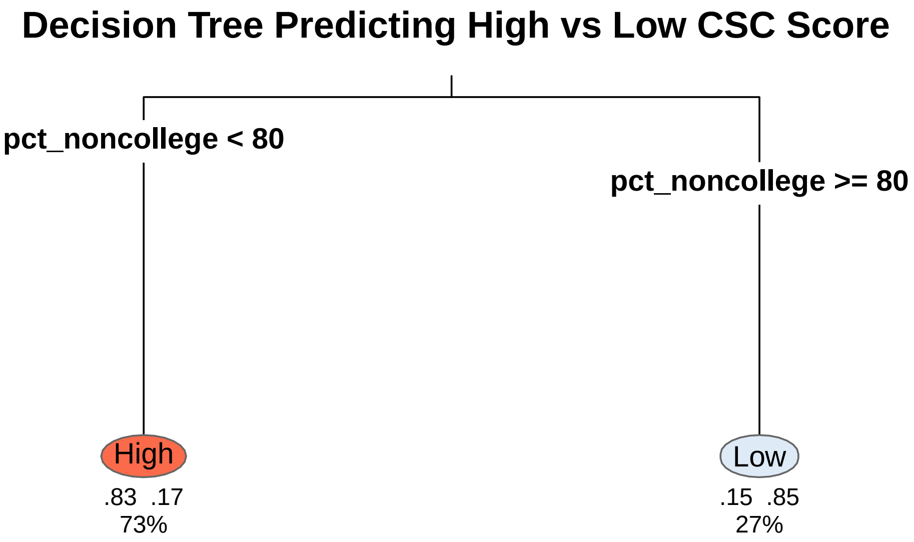
5 5. Advanced Analysis - Machine Learning

This decision tree analysis aims to identify the combinational patterns of socioeconomic factors that distinguish High versus Low CSC scores. The model evaluates all four variables simultaneously and automatically determines which combinations of conditions are associated with ZIP codes classified as having higher CSC levels.

```r
# Create binary outcome for classification
# High CSC (>= 3)  → "High"
# Low CSC (< 3)    → "Low"
zip_summary$CSC_binary <- ifelse(zip_summary$avg_CSC >= 3, "High", "Low")

# Fit Decision Tree Model
# The predictors are four socioeconomic indicators.
# The outcome variable is the binary CSC category.
tree <- rpart(
  CSC_binary ~ poverty_rate + pct_nonwhite + pct_non_english + pct_noncollege,
  data = zip_summary,
  method = "class"
)

# Plot the decision tree
# Visualization settings:
# - type = 3: show split labels and terminal node values
# - extra = 104: display class probabilities and class predictions
# - fallen.leaves = TRUE: tidy layout with leaves at the bottom
# - box.palette = "RdBu": two-color gradient
rpart.plot(
  tree,
  type = 3,
  extra = 104,
  under = TRUE,
  fallen.leaves = TRUE,
  roundint = FALSE,
  clip.right.labs = FALSE,
  box.palette = "RdBu",
  border.col = "gray40",
  cex = 0.9,
  main = "Decision Tree Predicting High vs Low CSC Score"
)
```



The decision tree shows that CSC levels are most clearly separated by a single factor: the percentage of adults without a college degree. ZIP codes where less than 80% of residents lack a college degree were classified as High CSC with an 83% probability, whereas ZIPs at or above 80% were classified as Low CSC with an 85% probability. This aligns with earlier findings that socioeconomic variables are generally weak predictors of CSC scores, though educational attainment appears to show a modest association at the ZIP level.

6 6. Conclusion

This analysis examined the spatial distribution of arts education levels (CSC scores) across Chicago ZIP codes and explored the socioeconomic factors that may shape these patterns. The spatial visualization revealed clear geographic disparities: higher CSC levels were concentrated in the North and central areas of the city, while lower levels appeared more frequently in the South and West Sides. However, regression analysis and interactive dashboard exploration showed that individual socioeconomic indicators, such as poverty rate, racial composition, non-English-speaking households, and educational attainment, displayed only weak associations with CSC scores, suggesting limited explanatory power.

The machine learning decision tree analysis reinforced these findings. Although the model evaluated all variables simultaneously, the percentage of adults without a college degree emerged as the single most informative factor for distinguishing High versus Low CSC ZIP codes. Additional variables or combinations did not improve classification accuracy. This suggests that socioeconomic indicators at the ZIP level alone do not strongly predict arts education quality, and that broader structural, institutional, or community-level factors may play a more significant role.

Overall, this study highlights that arts education equity cannot be fully explained by simple socioeconomic metrics, underscoring the complexity of the issue. From a policy perspective, this implies that interventions should go beyond demographic or economic characteristics and instead focus on strengthening community partnerships, improving program access, and supporting culturally and linguistically diverse learning environments. Future research may incorporate school-level operational data or partnership network information to develop more nuanced and predictive models.