DASC32103Project1-WIlliamBuckey

2025-02-05

```r
library(grid)
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(ggplot2)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```r
library(ggpubr)
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.3.2
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggpubr':
##
##     get_legend
```

```r
library(gridExtra)
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```r
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```r
library(dplyr)
```

```
##
```

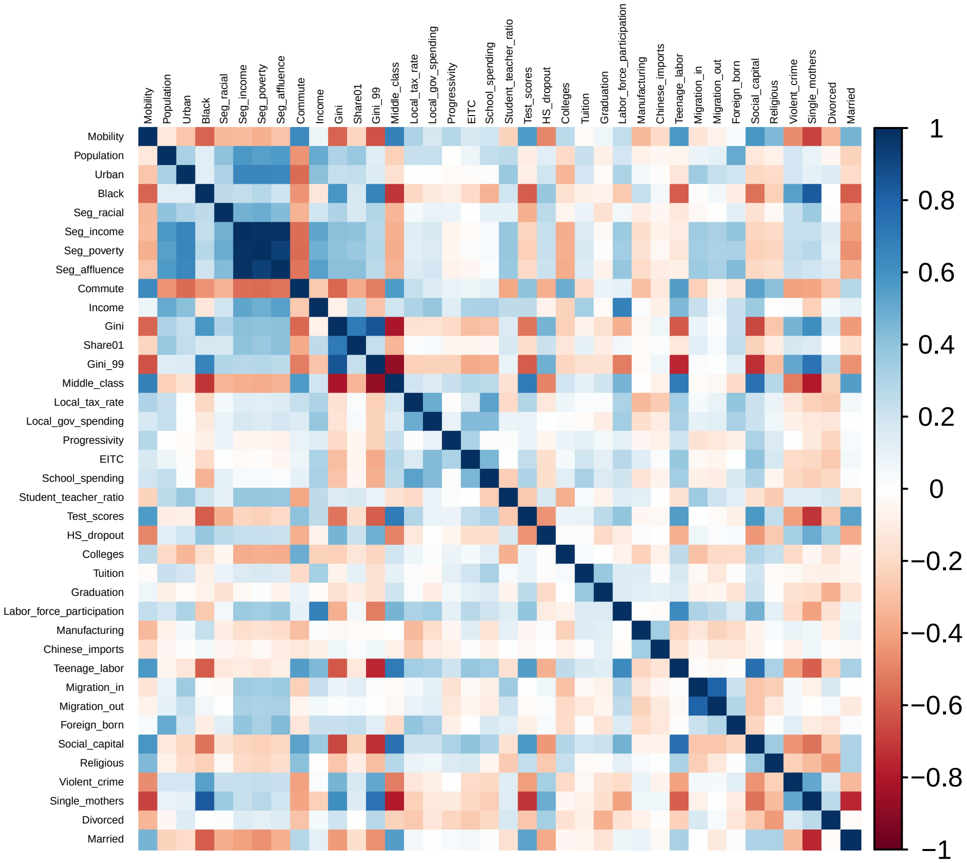
```
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
mobility_data <- read.csv("mobility-all.csv")
```

```
##                  Mobility                Population                     Urban
##                        12                         0                         0
##                     Black                Seg_racial                Seg_income
##                         0                         0                         0
##               Seg_poverty             Seg_affluence                   Commute
##                         0                         0                         0
##                    Income                      Gini                   Share01
##                         0                         0                        32
##                   Gini_99              Middle_class            Local_tax_rate
##                        32                        32                         1
##        Local_gov_spending             Progressivity                      EITC
##                         2                         0                         0
##           School_spending     Student_teacher_ratio               Test_scores
##                        10                        30                        36
##                HS_dropout                  Colleges                   Tuition
##                       148                       157                       161
##                Graduation Labor_force_participation             Manufacturing
##                       160                         0                         0
##           Chinese_imports             Teenage_labor              Migration_in
##                        19                        32                        17
##             Migration_out              Foreign_born            Social_capital
##                        17                         0                        19
##                 Religious             Violent_crime            Single_mothers
##                         0                        27                         0
##                  Divorced                   Married
##                         0                         0
```

```
# drop na values
mobility_data <- drop_na(mobility_data)
library(dplyr)

# correlation matrix
cor_matrix <- cor(mobility_data, use = "pairwise.complete.obs")
```

```
# heatmap
corrplot(cor_matrix,
         method = "color",
         tl.col = "black",
         tl.cex = 0.3)
```



```
cor_df <- as.data.frame(as.table(cor_matrix))

# Remove diagonal correlations
cor_df <- cor_df %>%
  filter(Var1 != Var2)

# Standardize Var1 & Var2
cor_df <- cor_df %>%
  dplyr::rowwise() %>%
  dplyr::mutate(pair = paste(sort(c(Var1, Var2)), collapse = "_")) %>%
  dplyr::distinct(pair, .keep_all = TRUE) %>%
  dplyr::select(-pair)

# Sort by correlation
top_corr <- cor_df %>%
  arrange(desc(abs(Freq))) %>%
  head(50)

# Print
print(top_corr)
```

```
## # A tibble: 50 x 3
## # Rowwise:
##    Var1           Var2            Freq
##    <fct>          <fct>          <dbl>
##  1 Seg_affluence  Seg_income     0.986
##  2 Seg_poverty    Seg_income     0.981
##  3 Seg_affluence  Seg_poverty    0.939
##  4 Middle_class   Gini_99       -0.870
##  5 Gini_99        Gini           0.857
##  6 Single_mothers Black          0.837
##  7 Middle_class   Gini          -0.815
##  8 Migration_out  Migration_in   0.804
##  9 Single_mothers Middle_class  -0.791
## 10 Social_capital Teenage_labor  0.760
## # i 40 more rows
```

```r
library(dplyr)
# Define policy-driven variables
policy_vars <- c("Local_tax_rate", "Local_gov_spending", "Progressivity", "Gini",
                 "School_spending", "Gini_99", "Test_scores",
                 "HS_dropout", "Middle_class", "Social_capital",
                 "Colleges", "Tuition", "Single_mothers")

# Correlation matrix
cor_matrix <- cor(mobility_data, use = "pairwise.complete.obs")

# Convert matrix into a dataframe
cor_df <- as.data.frame(as.table(cor_matrix))

# Remove diagonal correlations
cor_df <- cor_df %>%
  filter(Var1 != Var2)

# Standardize Var1 & Var2
cor_df <- cor_df %>%
  dplyr::rowwise() %>%
  dplyr::mutate(pair = paste(sort(c(Var1, Var2)), collapse = "_")) %>%
  dplyr::distinct(pair, .keep_all = TRUE) %>%
  dplyr::select(-pair)

# Find top 5 correlated variables
top_correlations <- list()

for (var in policy_vars) {
  top_5 <- cor_df %>%
    filter(Var1 == var | Var2 == var) %>%
    arrange(desc(abs(Freq))) %>%
    head(5)
  top_correlations[[var]] <- top_5
}

# Display
print(top_correlations)
```

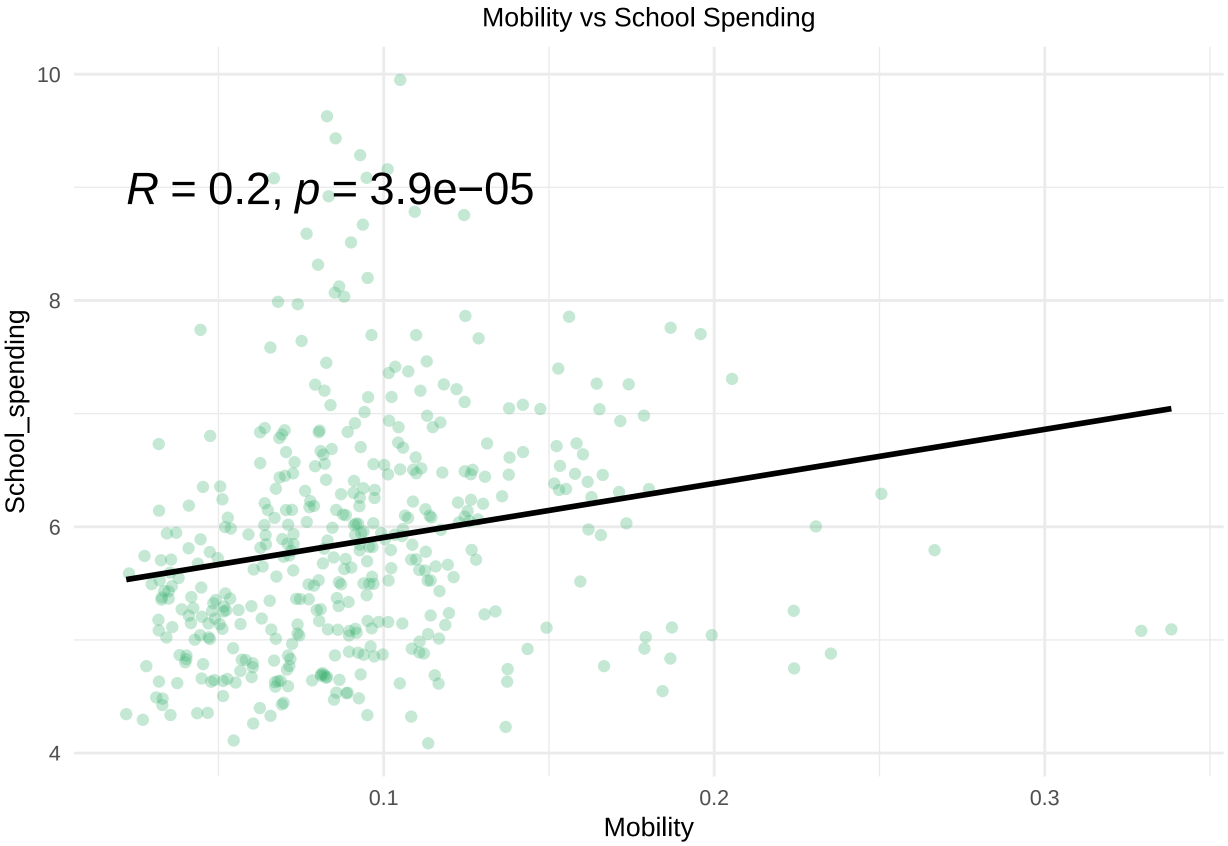
```
## $Local_tax_rate
## # A tibble: 5 x 3
## # Rowwise:
##   Var1              Var2            Freq
##   <fct>             <fct>          <dbl>
## 1 School_spending   Local_tax_rate  0.538
## 2 Local_gov_spending Local_tax_rate  0.496
## 3 Foreign_born      Local_tax_rate  0.399
## 4 Teenage_labor     Local_tax_rate  0.344
## 5 Manufacturing     Local_tax_rate -0.327
##
## $Local_gov_spending
## # A tibble: 5 x 3
## # Rowwise:
##   Var1                     Var2                Freq
##   <fct>                    <fct>              <dbl>
## 1 Local_gov_spending       Local_tax_rate     0.496
## 2 School_spending          Local_gov_spending 0.435
## 3 EITC                     Local_gov_spending 0.430
## 4 Local_gov_spending       Income             0.383
## 5 Labor_force_participation Local_gov_spending 0.345
##
## $Progressivity
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2           Freq
##   <fct>          <fct>         <dbl>
## 1 Social_capital Progressivity  0.316
## 2 EITC           Progressivity  0.312
## 3 Progressivity  Mobility       0.286
## 4 Progressivity  Gini_99       -0.222
## 5 Progressivity  Middle_class   0.221
##
## $Gini
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2    Freq
##   <fct>          <fct>  <dbl>
## 1 Gini_99        Gini   0.857
## 2 Middle_class   Gini  -0.815
## 3 Share01        Gini   0.701
## 4 Social_capital Gini  -0.662
## 5 Teenage_labor  Gini  -0.618
##
## $School_spending
## # A tibble: 5 x 3
## # Rowwise:
##   Var1            Var2                Freq
##   <fct>           <fct>              <dbl>
## 1 School_spending Local_tax_rate      0.538
## 2 School_spending EITC                0.452
## 3 School_spending Local_gov_spending  0.435
## 4 School_spending Gini_99            -0.354
## 5 Teenage_labor   School_spending     0.344
##
```

```
## $Gini_99
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2      Freq
##   <fct>          <fct>    <dbl>
## 1 Middle_class   Gini_99 -0.870
## 2 Gini_99        Gini     0.857
## 3 Teenage_labor  Gini_99 -0.750
## 4 Social_capital Gini_99 -0.737
## 5 Single_mothers Gini_99  0.734
##
## $Test_scores
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2           Freq
##   <fct>          <fct>         <dbl>
## 1 Single_mothers Test_scores  -0.718
## 2 Test_scores    Middle_class  0.709
## 3 Test_scores    Gini_99      -0.606
## 4 Test_scores    Black        -0.600
## 5 Social_capital Test_scores   0.576
##
## $HS_dropout
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2           Freq
##   <fct>          <fct>         <dbl>
## 1 Single_mothers HS_dropout    0.494
## 2 HS_dropout     Middle_class -0.490
## 3 HS_dropout     Mobility     -0.481
## 4 HS_dropout     Gini_99       0.480
## 5 HS_dropout     Gini          0.468
##
## $Middle_class
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2           Freq
##   <fct>          <fct>         <dbl>
## 1 Middle_class   Gini_99      -0.870
## 2 Middle_class   Gini         -0.815
## 3 Single_mothers Middle_class -0.791
## 4 Social_capital Middle_class  0.740
## 5 Middle_class   Black        -0.712
##
## $Social_capital
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2            Freq
##   <fct>          <fct>          <dbl>
## 1 Social_capital Teenage_labor  0.760
## 2 Social_capital Middle_class   0.740
## 3 Social_capital Gini_99       -0.737
## 4 Social_capital Gini          -0.662
## 5 Social_capital Mobility       0.585
```

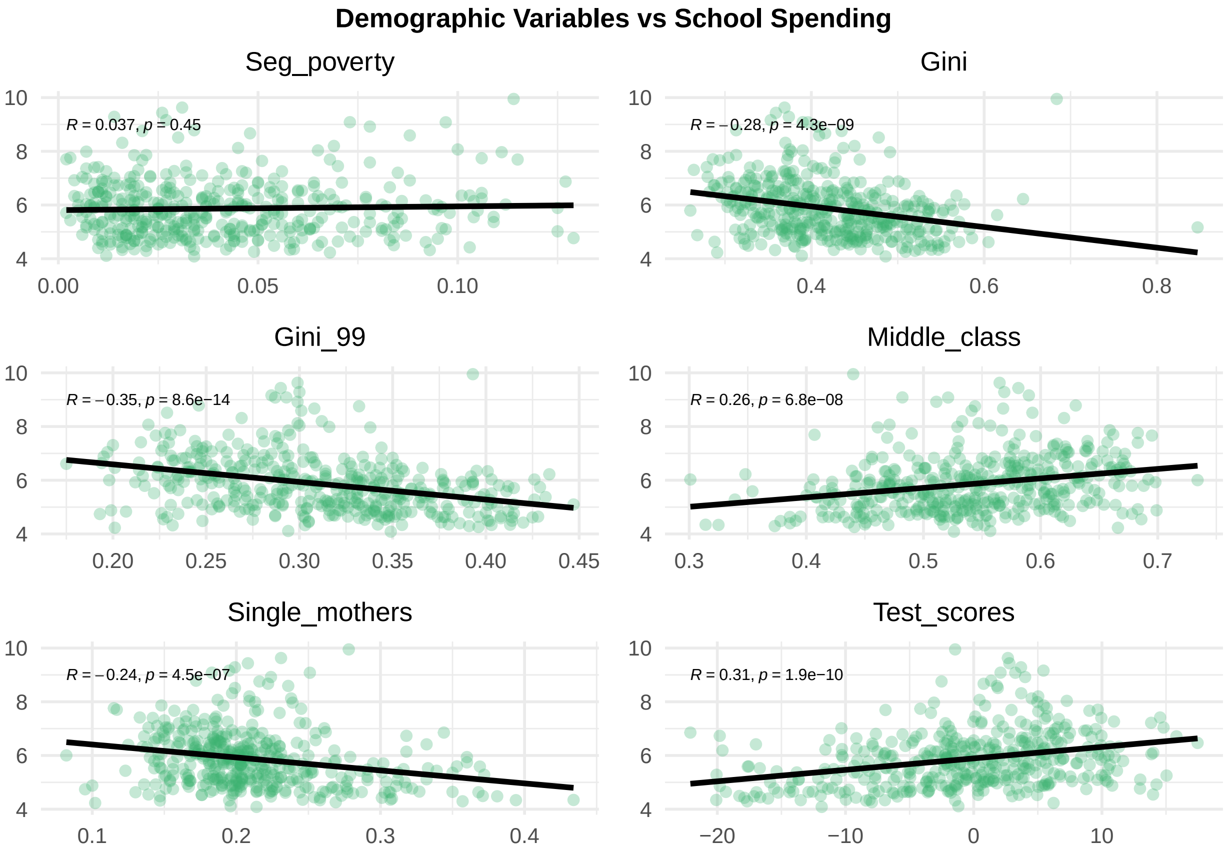
```
##
## $Colleges
## # A tibble: 5 x 3
## # Rowwise:
##   Var1     Var2                   Freq
##   <fct>    <fct>                 <dbl>
## 1 Colleges Commute               0.499
## 2 Colleges Seg_affluence        -0.362
## 3 Colleges Seg_income           -0.361
## 4 Colleges Seg_poverty          -0.356
## 5 Colleges Student_teacher_ratio -0.351
##
## $Tuition
## # A tibble: 5 x 3
## # Rowwise:
##   Var1       Var2              Freq
##   <fct>      <fct>            <dbl>
## 1 Graduation Tuition          0.373
## 2 Tuition    Income           0.330
## 3 Tuition    School_spending  0.324
## 4 Tuition    Population       0.222
## 5 Tuition    Commute         -0.198
##
## $Single_mothers
## # A tibble: 5 x 3
## # Rowwise:
##   Var1           Var2             Freq
##   <fct>          <fct>           <dbl>
## 1 Single_mothers Black           0.837
## 2 Single_mothers Middle_class   -0.791
## 3 Married        Single_mothers -0.753
## 4 Single_mothers Gini_99         0.734
## 5 Single_mothers Test_scores    -0.718
```

```r
# Create scatter plots
plot_scatter <- function(x_var, color, text_size = 5, r_p_size = 5, keep_axis_titles = FALSE) {
  ggplot(mobility_data, aes(.data[[x_var]], School_spending)) +
    geom_point(color = color, alpha = .3) +
    geom_smooth(method = "lm", color = "black", se = FALSE) +  # Add linear regression line
    stat_cor(label.x = min(mobility_data[[x_var]], na.rm = TRUE),
             label.y = max(mobility_data$School_spending, na.rm = TRUE) * 0.9,
             size = r_p_size) +  # Adds R & p-values
    ggtitle(paste(x_var)) +
    theme_minimal() +
    theme(
      axis.title = if (keep_axis_titles) element_text(size = 10) else element_blank(),
      axis.text = element_text(size = 8),
      plot.title = element_text(hjust = 0.5, size = 10)
    )
}

# Create plot vs mobility
mobility_plot <- plot_scatter("Mobility", "mediumseagreen", text_size = 5, r_p_size = 6, keep_axis_titl
  ggtitle("Mobility vs School Spending")
```

```
# Create plots for other variables
p1 <- plot_scatter("Seg_poverty", "mediumseagreen", r_p_size = 2)
p2 <- plot_scatter("Gini", "mediumseagreen", r_p_size = 2)
p3 <- plot_scatter("Gini_99", "mediumseagreen", r_p_size = 2)
p4 <- plot_scatter("Middle_class", "mediumseagreen", r_p_size = 2)
p5 <- plot_scatter("Single_mothers", "mediumseagreen", r_p_size = 2)
p6 <- plot_scatter("Test_scores", "mediumseagreen", r_p_size = 2)

# Display
print(mobility_plot)
```

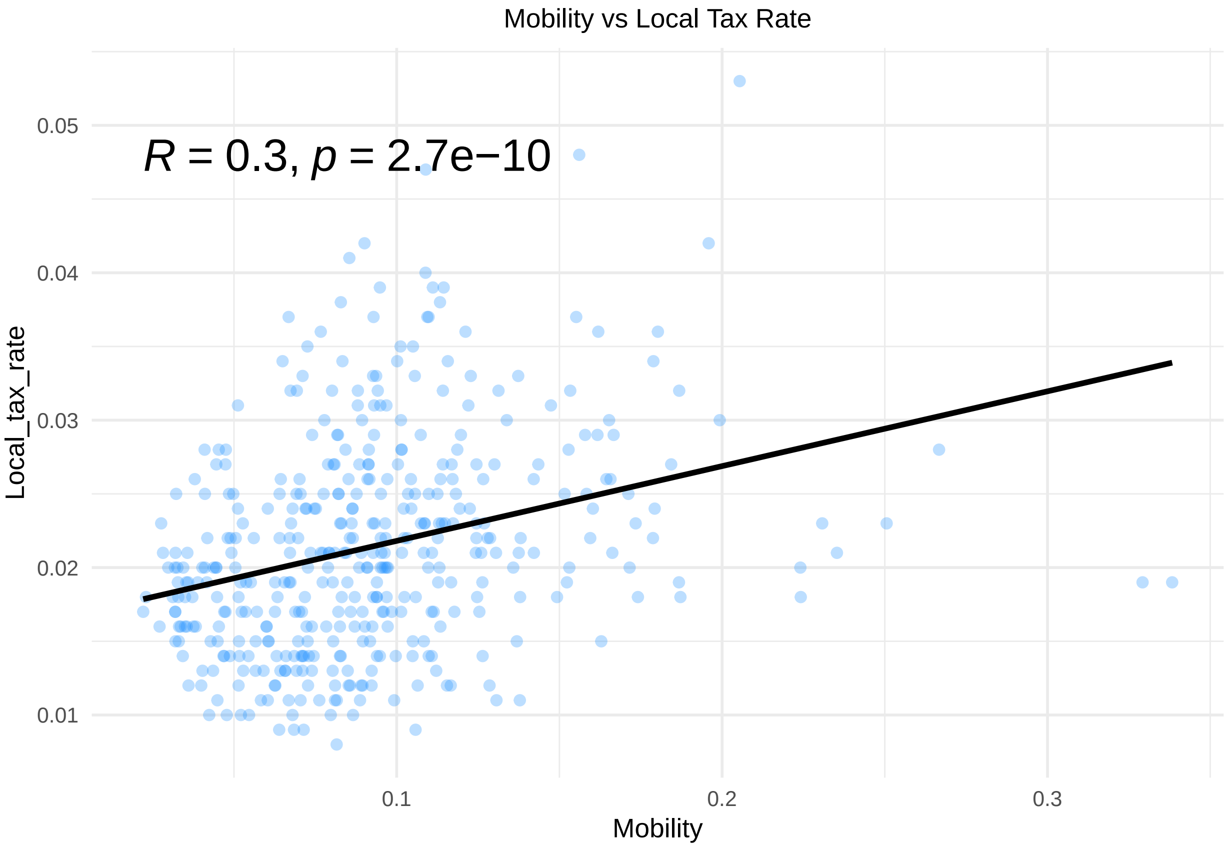


```
# Display all other plots on one page
grid.arrange(
  arrangeGrob(p1, p2, p3, p4, p5, p6, ncol = 2,
              top = textGrob("Demographic Variables vs School Spending",
                             gp = gpar(fontsize = 10, fontface = "bold")))
)
```

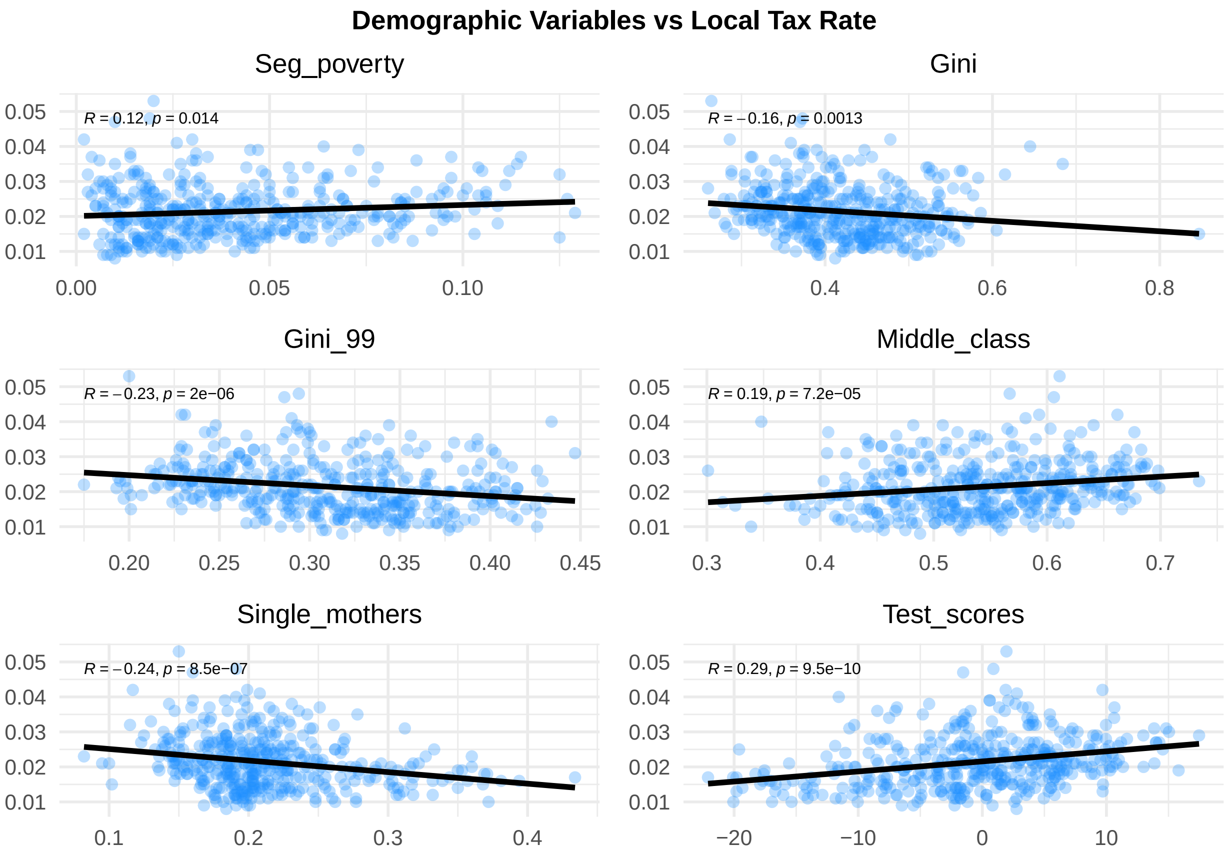


```r
data <- mobility_data
# Create scatter plots
plot_scatter <- function(x_var, color, text_size = 5, r_p_size = 5, keep_axis_titles = FALSE) {
  ggplot(data, aes(.data[[x_var]], Local_tax_rate)) +
    geom_point(color = color, alpha = .3) +
    geom_smooth(method = "lm", color = "black", se = FALSE) +
    stat_cor(label.x = min(data[[x_var]]),
             label.y = max(data$Local_tax_rate, na.rm = TRUE) * 0.9,
             size = r_p_size) +
    ggtitle(paste(x_var)) +
    theme_minimal() +
    theme(
      axis.title = if (keep_axis_titles) element_text(size = 10) else element_blank(),
      axis.text = element_text(size = 8),
      plot.title = element_text(hjust = 0.5, size = 10)
    )
}

# Create plot vs Mobility
mobility_plot <- plot_scatter("Mobility", "dodgerblue", text_size = 5, r_p_size = 6, keep_axis_titles =
  ggtitle("Mobility vs Local Tax Rate")

# Create plots for other variables
p1 <- plot_scatter("Seg_poverty", "dodgerblue", r_p_size = 2)
p2 <- plot_scatter("Gini", "dodgerblue", r_p_size = 2)
p3 <- plot_scatter("Gini_99", "dodgerblue", r_p_size = 2)
p4 <- plot_scatter("Middle_class", "dodgerblue", r_p_size = 2)
```

```r
p5 <- plot_scatter("Single_mothers", "dodgerblue", r_p_size = 2)
p6 <- plot_scatter("Test_scores", "dodgerblue", r_p_size = 2)

# Display
print(mobility_plot)
```

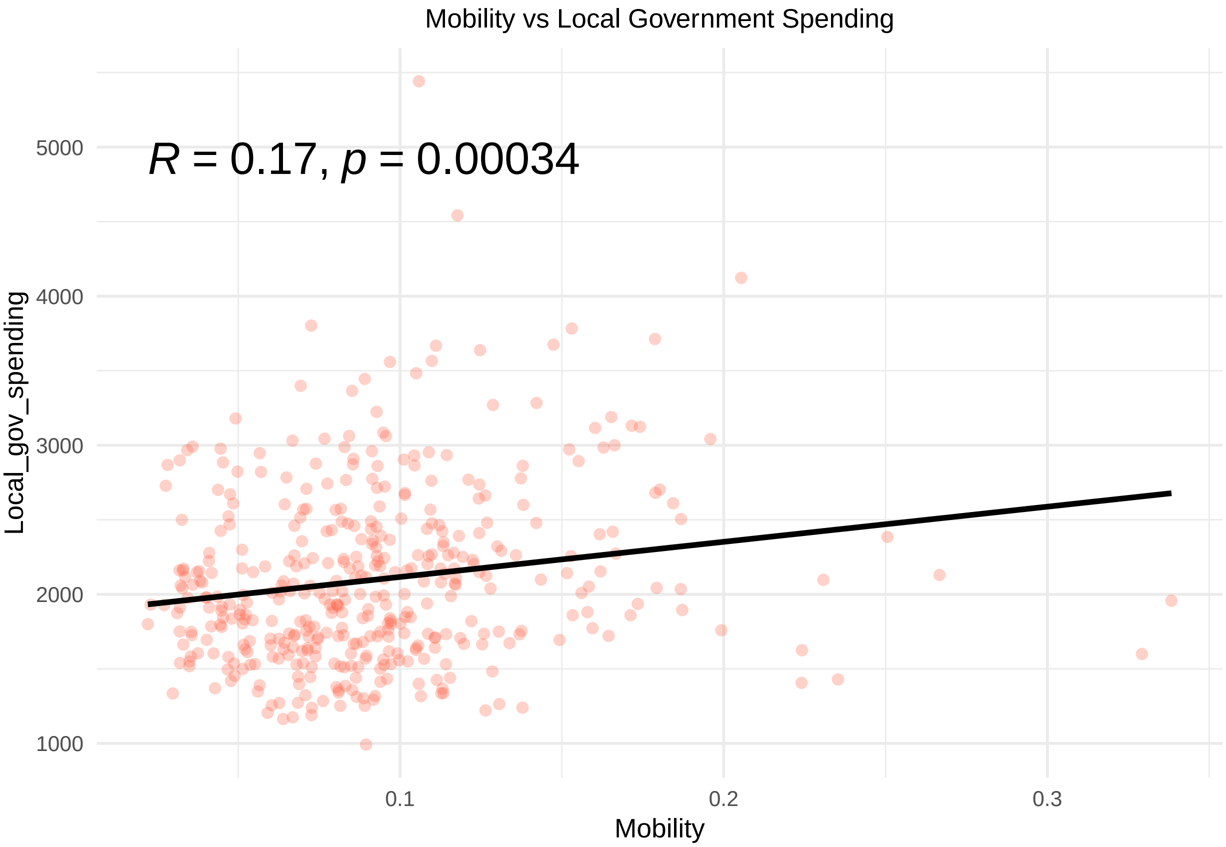


```r
# Arrange and display all other plots on one page
grid.arrange(
  arrangeGrob(p1, p2, p3, p4, p5, p6, ncol = 2,
              top = textGrob("Demographic Variables vs Local Tax Rate",
                             gp = gpar(fontsize = 10, fontface = "bold")))
)
```

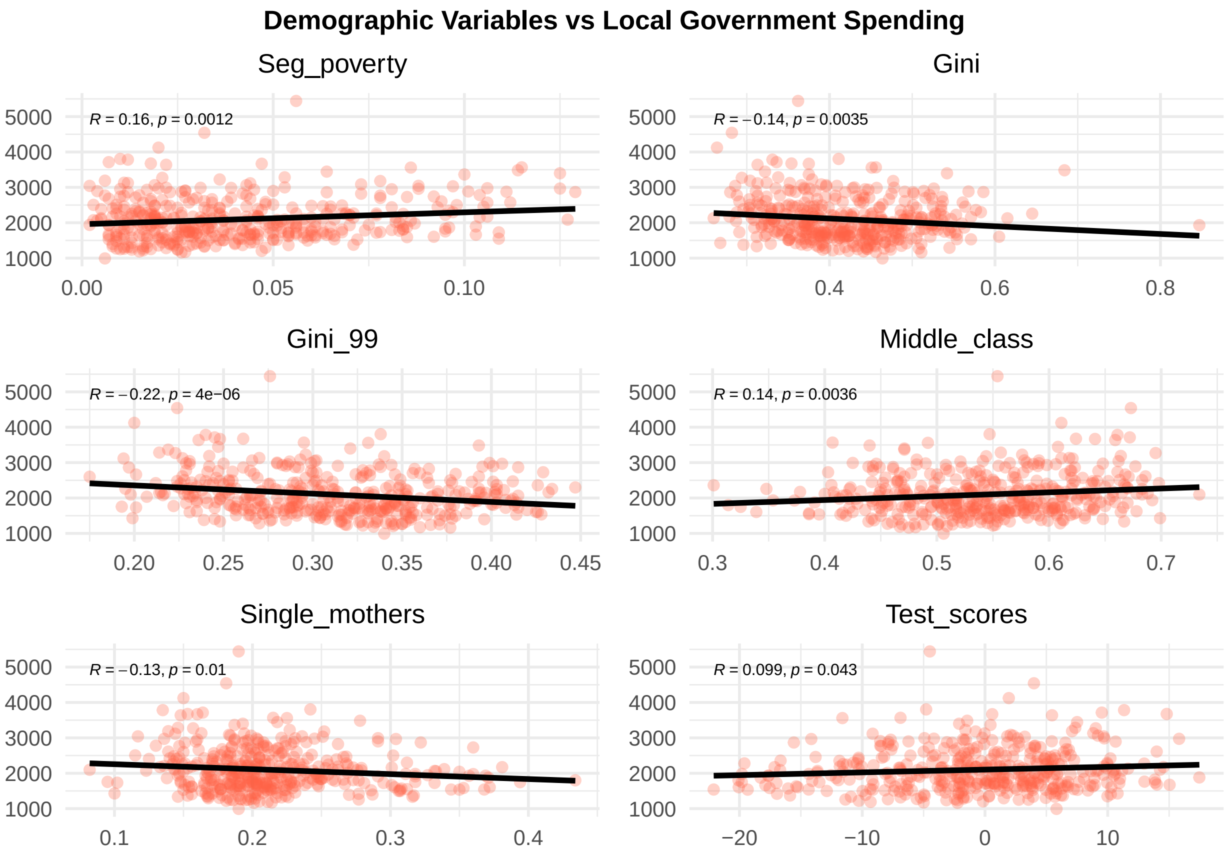


```
# Create scatter plots
plot_scatter <- function(x_var, color, text_size = 5, r_p_size = 5, keep_axis_titles = FALSE) {
  ggplot(data, aes(.data[[x_var]], Local_gov_spending)) +
    geom_point(color = color, alpha = .3, na.rm = TRUE) +
    geom_smooth(method = "lm", color = "black", se = FALSE) +
    stat_cor(label.x = min(data[[x_var]], na.rm = TRUE),
             label.y = max(data$Local_gov_spending, na.rm = TRUE) * 0.9,
             size = r_p_size) +
    ggtitle(paste(x_var)) +
    theme_minimal() +
    theme(
      axis.title = if (keep_axis_titles) element_text(size = 10) else element_blank(),
      axis.text = element_text(size = 8),
      plot.title = element_text(hjust = 0.5, size = 10)
    )
}

# Create plot vs Mobility
mobility_plot <- plot_scatter("Mobility", "tomato", text_size = 5, r_p_size = 6, keep_axis_titles = TRU
  ggtitle("Mobility vs Local Government Spending")

# Create plots for other variables
p1 <- plot_scatter("Seg_poverty", "tomato", r_p_size = 2)
p2 <- plot_scatter("Gini", "tomato", r_p_size = 2)
p3 <- plot_scatter("Gini_99", "tomato", r_p_size = 2)
p4 <- plot_scatter("Middle_class", "tomato", r_p_size = 2)
p5 <- plot_scatter("Single_mothers", "tomato", r_p_size = 2)
```

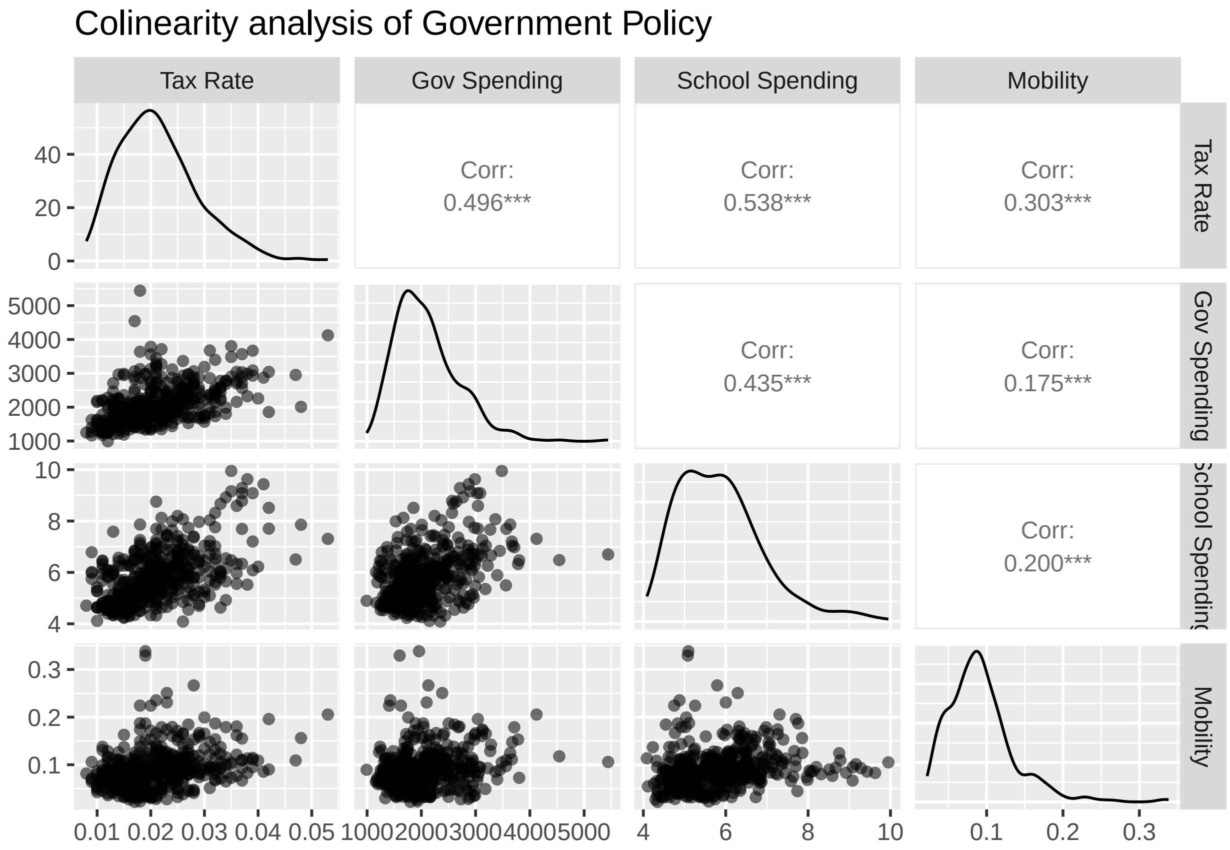
```
p6 <- plot_scatter("Test_scores", "tomato", r_p_size = 2)

# Display
print(mobility_plot)
```



```
# Arrange and display all other plots on one page
grid.arrange(
  arrangeGrob(p1, p2, p3, p4, p5, p6, ncol = 2,
              top = textGrob("Demographic Variables vs Local Government Spending",
                             gp = gpar(fontsize = 10, fontface = "bold")))
)
```



```
mobility_data[c("Local_tax_rate", "Local_gov_spending", "School_spending", "Mobility")] %>%
  ggpairs(aes(alpha = 0.5),
          upper = list(continuous = wrap("cor", size = 3)),
          columnLabels = c("Tax Rate", "Gov Spending", "School Spending", "Mobility"),
          title = "Colinearity analysis of Government Policy",
          progress = FALSE)
```



```
# Define predictor variables
candidate_vars <- c("Local_tax_rate", "Local_gov_spending", "School_spending",
                    "Test_scores", "Single_mothers", "Seg_poverty", "Gini_99", "Gini", "Middle_class")

# Model formula
full_formula <- as.formula(paste("Mobility ~", paste(candidate_vars, collapse = " + ")))

# Fit the full model
full_model <- lm(full_formula, data = mobility_data)
summary(full_model)
```

```
##
## Call:
## lm(formula = full_formula, data = mobility_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062298 -0.017234 -0.004222  0.011418  0.209359
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.593e-01  4.601e-02   3.462 0.000592 ***
## Local_tax_rate     1.213e+00  2.543e-01   4.770 2.57e-06 ***
## Local_gov_spending  3.579e-06  2.944e-06   1.216 0.224763
## School_spending   -5.169e-03  1.751e-03  -2.952 0.003335 **
## Test_scores        4.388e-04  3.031e-04   1.448 0.148414
## Single_mothers    -2.419e-01  4.926e-02  -4.910 1.32e-06 ***
## Seg_poverty        -2.719e-01  5.947e-02  -4.572 6.42e-06 ***
## Gini_99            -1.375e-01  6.519e-02  -2.109 0.035570 *
## Gini               -1.466e-02  3.994e-02  -0.367 0.713723
## Middle_class        7.614e-02  4.760e-02   1.600 0.110483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02943 on 408 degrees of freedom
## Multiple R-squared:  0.5638, Adjusted R-squared:  0.5542
## F-statistic:  58.6 on 9 and 408 DF,  p-value: < 2.2e-16
```

```
# Stepwise selection (both directions)
best_model <- stepAIC(full_model, direction = "both", trace = FALSE)
summary(best_model)
```

```
##
## Call:
## lm(formula = Mobility ~ Local_tax_rate + School_spending + Single_mothers +
##     Seg_poverty + Gini_99 + Middle_class, data = mobility_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.060242 -0.017909 -0.004178  0.011566  0.211910
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.146651   0.042840   3.423 0.000681 ***
## Local_tax_rate   1.356093   0.236871   5.725 2.00e-08 ***
## School_spending -0.004466   0.001703  -2.623 0.009048 **
## Single_mothers  -0.260381   0.045190  -5.762 1.63e-08 ***
## Seg_poverty     -0.269065   0.056550  -4.758 2.71e-06 ***
## Gini_99         -0.147449   0.053949  -2.733 0.006544 **
## Middle_class     0.100953   0.043411   2.326 0.020529 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02944 on 411 degrees of freedom
## Multiple R-squared:  0.5603, Adjusted R-squared:  0.5539
## F-statistic:  87.3 on 6 and 411 DF,  p-value: < 2.2e-16
```

```
# Multicollinearity check using VIF
vif_values <- vif(best_model)
print(vif_values)
```

```
##  Local_tax_rate School_spending  Single_mothers     Seg_poverty         Gini_99
##        1.474320        1.551267        2.789308        1.205282        4.539181
##    Middle_class
##        5.496419
```

```
# Final model
final_model <- best_model
summary(final_model)
```

```
##
## Call:
## lm(formula = Mobility ~ Local_tax_rate + School_spending + Single_mothers +
##     Seg_poverty + Gini_99 + Middle_class, data = mobility_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.060242 -0.017909 -0.004178  0.011566  0.211910
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.146651   0.042840   3.423 0.000681 ***
## Local_tax_rate   1.356093   0.236871   5.725 2.00e-08 ***
## School_spending -0.004466   0.001703  -2.623 0.009048 **
## Single_mothers  -0.260381   0.045190  -5.762 1.63e-08 ***
## Seg_poverty     -0.269065   0.056550  -4.758 2.71e-06 ***
## Gini_99         -0.147449   0.053949  -2.733 0.006544 **
## Middle_class     0.100953   0.043411   2.326 0.020529 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02944 on 411 degrees of freedom
## Multiple R-squared:  0.5603, Adjusted R-squared:  0.5539
## F-statistic:  87.3 on 6 and 411 DF,  p-value: < 2.2e-16
```

```r
# Final model's formula to a GLM with Gaussian family
model_glm <- glm(formula(final_model), data = mobility_data, family = gaussian())

cv_error <- cv.glm(mobility_data, model_glm, K = 10)
cat("Cross-Validation Error:", cv_error$delta[1], "\n")
```

```
## Cross-Validation Error: 0.0008873957
```