Project 1

Ruth Walters

2025-02-19

Introduction

Economic mobility, or the ability of an individual to raise their economic status throughout their lifetime, is a marker of a healthy society. As economic mobility declines and income inequality rises throughout the United States, it is of increasing interest to determine which factors contribute to immobility. In this paper, we will investigate the correlation between economic, educational, and policy factors that contribute to economic mobility. We hypothesize that economic factors such as income inequality, will be most predictive of economic mobility.

Exploratory data analysis

```
# View NAs
nas <- colSums(is.na(mobility))
print(nas[nas > 0])
```

```
##              Mobility               Share01               Gini_99
##                    12                    32                    32
##          Middle_class       Local_tax_rate   Local_gov_spending
##                    32                     1                     2
##       School_spending Student_teacher_ratio          Test_scores
##                    10                    30                    36
##            HS_dropout              Colleges               Tuition
##                   148                   157                   161
##            Graduation      Chinese_imports        Teenage_labor
##                   160                    19                    32
##          Migration_in       Migration_out       Social_capital
##                    17                    17                    19
##         Violent_crime
##                    27
```

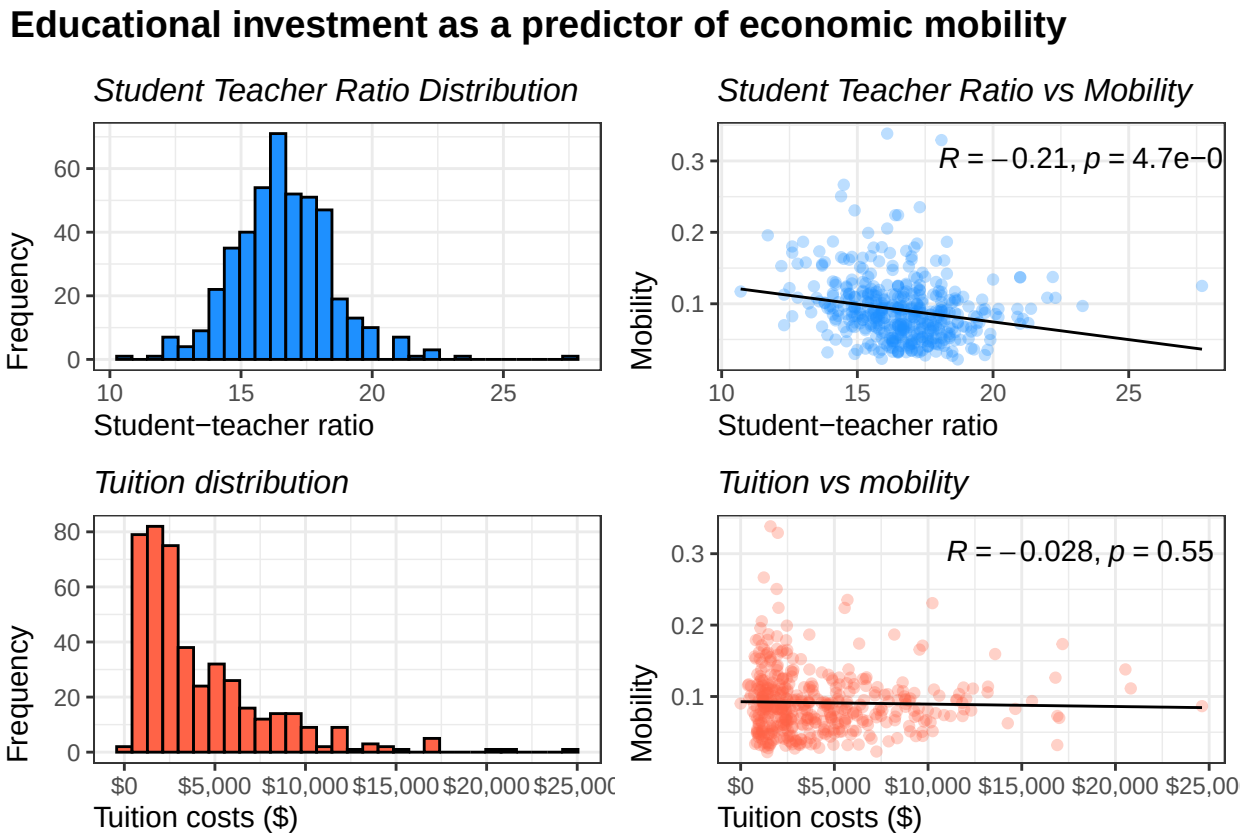
This dataset contains several rows for which one or more than one value is `NA`. Three steps were taken to eliminate `NA`s from the dataset while preserving its integrity.

1. *Drop the 12 rows that do not contain a value for* ***`Mobility`*** | These rows are useless for linear analysis because they do not contain the variable we are attempting to predict.
2. *Drop the features that contain a high incidence of* ***`NA`****s* | Any features that contained more than 100 NAs were designated as too poor in quality to be useful for the linear model. While some of these features were used for exploratory data analysis, they were removed from the dataset prior to modeling.
3. *Drop all remaining* ***`NA`*** *values* | After removing the most `NA` values, a small amount of rows with `NA`s remained. These rows were dropped.

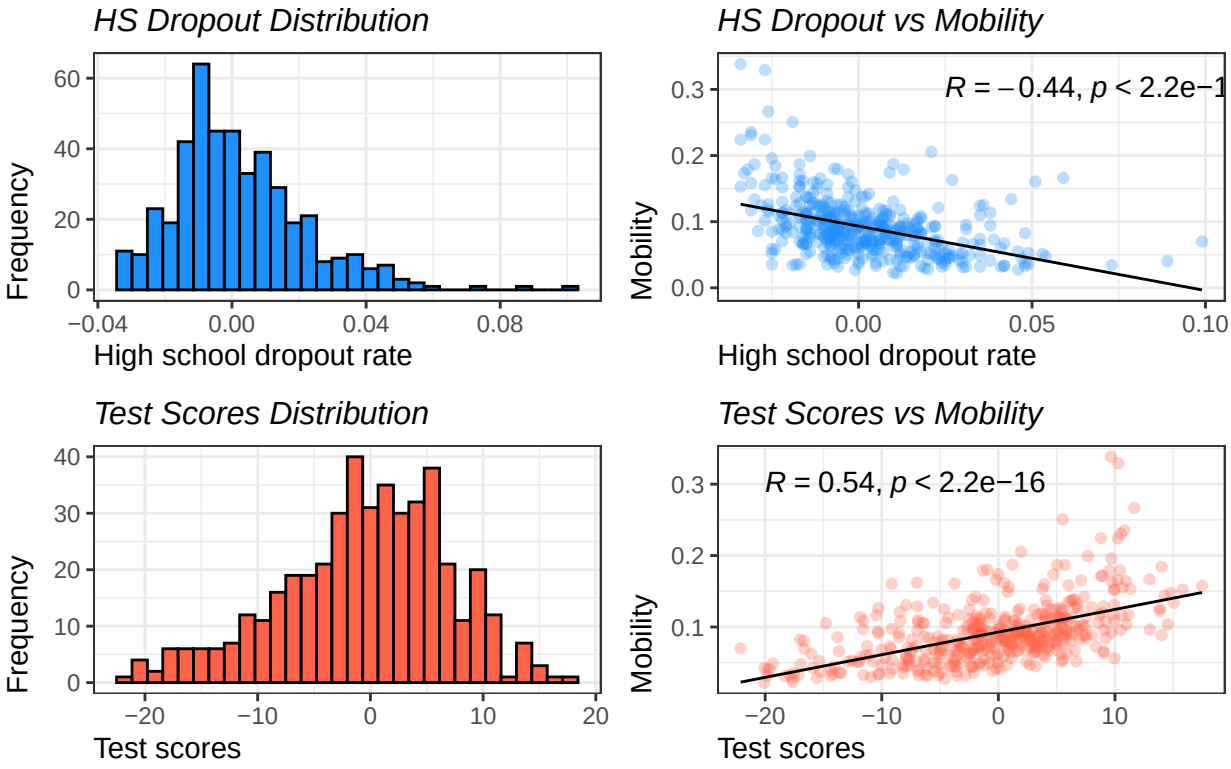
Simply dropping all rows with `NA`s would have resulted in a reduction of 39.4062078% of the data whereas our three-step procedure only resulted in a reduction of 14.5748988%.

Additionally, qualitative (non-numeric and non-quantitative) variables such as those representing latitude and longitude, state/region names, and the ID tag, were removed.

Education analysis



Educational outcomes as a predictor of economic mobility



Policy variables

When examining the mobility dataset, there seemed to be 2 main features that are directly impacted by government policies:

1. *Local tax rate* | Fraction of all income going to local taxes
2. *School expenditures* | Average spending per pupil in public schools
3. *Local government spending* | Local government spending per capita

Out of all the variables in the mobility dataset, the local tax rate and school expenditure are the most directly affected by government policy; that is, local tax rate and school spending are dependent on how the government chooses to raise and spend funds.

While `Chinese_imports` and `Manufacturing` were taken into consideration as policy-based predictors, both features were deemed to stem more from business practices that government policy. Chinese imports are impacted by US foreign policy, however, it would be difficult to untangle the competing market and regulatory factors,so it is difficult to say that these variables are linked to US government policy.

After examining the dataset to verify that `School_spending` and `Local_tax_rate` were valid, there was found to be very little missing values for these variables. In fact, it's less than mobility! This gives assurance that these two variables are valid enough within the dataset to be used as predicting variables.

Government policy is first examined by identifying the relationship between the three variables that are selected as predictors and verifying that these variables have a positive correlation between themselves. To verify these relations, a correlation matrix was created using pairwise relations. We highlight the policy-associated variables and display the top 5 variables that are correlated with each.

```
# Print strongest correlations
print(top_correlations$School_spending)
```

```
## # A tibble: 5 x 3
## # Rowwise:
##   Var1            Var2                Freq
##   <fct>           <fct>              <dbl>
## 1 School_spending Local_tax_rate      0.524
## 2 School_spending EITC                0.382
## 3 School_spending Local_gov_spending  0.367
## 4 Teenage_labor   School_spending     0.344
## 5 School_spending Gini_99            -0.325
```

```
print(top_correlations$Local_tax_rate)
```

```
## # A tibble: 5 x 3
## # Rowwise:
##   Var1               Var2             Freq
##   <fct>              <fct>           <dbl>
## 1 School_spending    Local_tax_rate  0.524
## 2 Local_gov_spending Local_tax_rate  0.450
## 3 Manufacturing      Local_tax_rate -0.373
## 4 Teenage_labor      Local_tax_rate  0.372
## 5 Local_tax_rate     Mobility        0.346
```

```
print(top_correlations$Local_gov_spending)
```

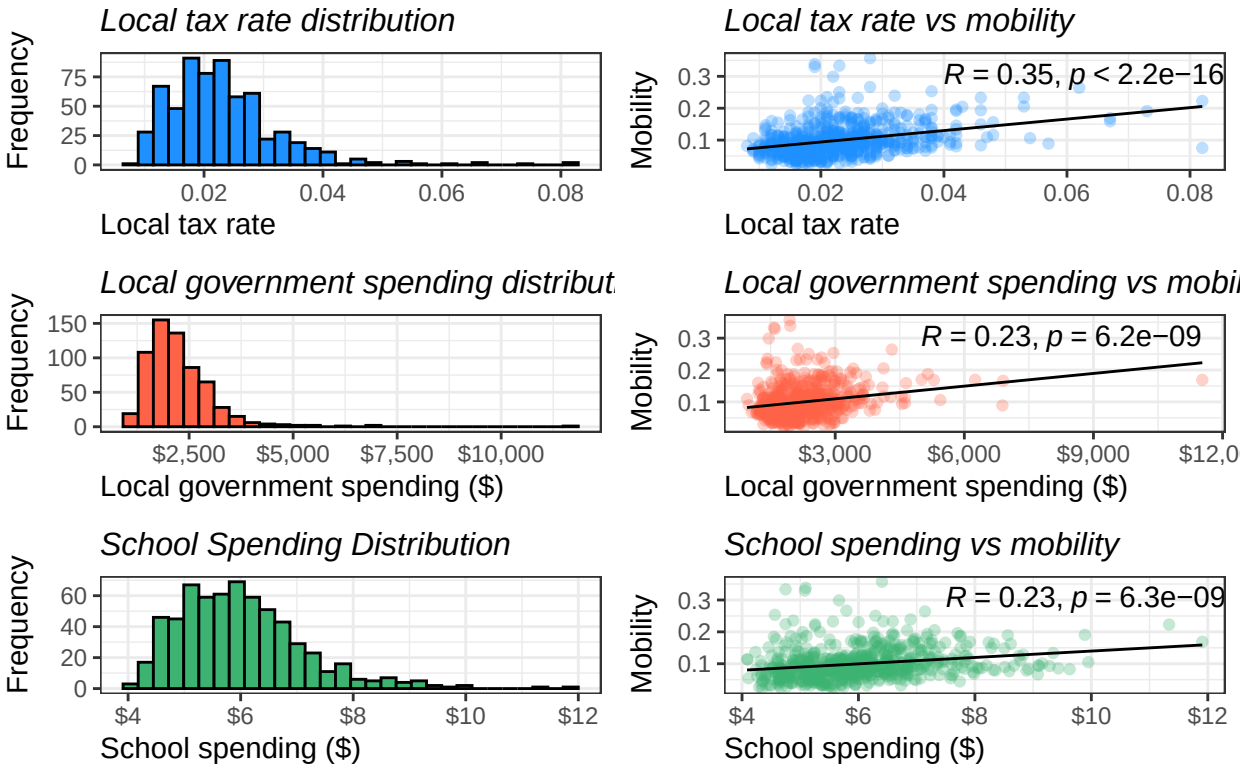
```
## # A tibble: 5 x 3
## # Rowwise:
##   Var1               Var2                Freq
##   <fct>              <fct>              <dbl>
## 1 Local_gov_spending Local_tax_rate     0.450
## 2 School_spending    Local_gov_spending 0.367
## 3 Local_gov_spending Income             0.301
## 4 Foreign_born       Local_gov_spending 0.295
## 5 Teenage_labor      Local_gov_spending 0.293
```

For the 3 predicting variables, all 3 are were correlated the highly with each other. `School_spending`, `Local_tax_rate`, and `Local_gov_spending` all have similar correlations, which verifies the covariate reliance on government policy.

From here, these 3 variables are used to examine the relations each has on demographic factors that directly affect an individual's economic mobility position. The demographic variables examined were `Seg_poverty`, `Gini`, `Gini_99`, `Middle_class`, `Single_mothers`, and `Test scores`. These were all used as sample variable measures because, based on previous findings, these variables all affect mobility. The sample variables are then used to generate scatter plots (see **Appendix A**) examining each specific predicting variable. A linear regression line and R and p values are displayed to elaborate on the linear trend lines.

`School_spending` seemed to be the only variable that has a linear relationship with the sample variables. While `Local_tax_rate` and `Local_gov_spending` did have linear relationships with the demographic variables, they were on the whole weak. Further, histograms and scatter plots that examine the distribution of each variable and its individual relationship to `Mobility` were generated to determine if the predictors have an initial linear regression relationship with mobility.

Policy predictors of mobility



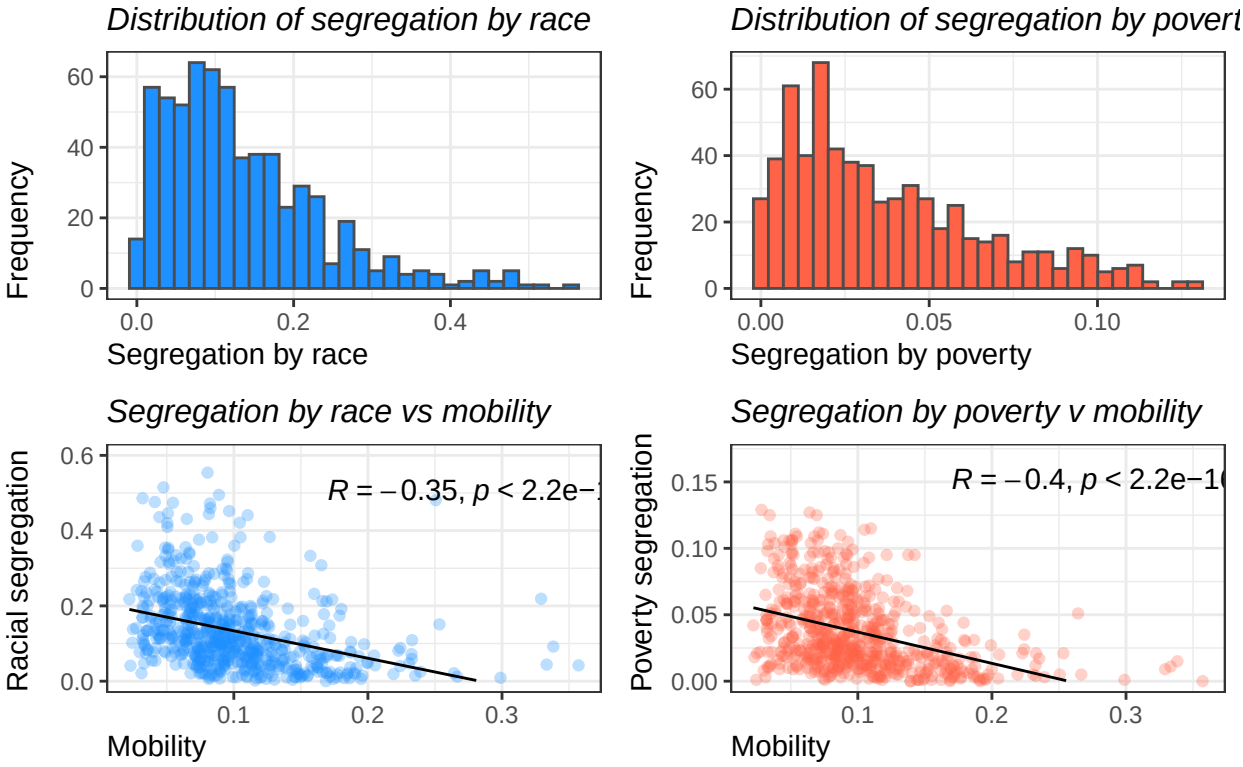
These results state to us that while there might be outlier data in all 3 relational graphs, there is still a somewhat linear relationship between mobility and tax rate. This is interesting because `School_spending` had was more highly correlated with `Mobility` than with demographic factors. `School_spending` and `Local_gov_spedning` do not seem to show the linear trends, but since `School_spending` had is so highly correlated with other demographic factors, it's inclusion in the model is justified.

Segregation factors

Covariate analysis of segregation factors (`Seg_racial`, `Seg_poverty`, `Seg_affluence` and `Seg_income`- see **Appendix B.4**) indicated that, while segregation on poverty lines is not particularly well correlated with segregation on racial lines, it is highly associated with segregation by affluence and segregation by income, which are also highly associated with each other. Since `Seg_poverty`, `Seg_affluence` and `Seg_income` are so strongly co-linear, `Seg_affluence` and `Seg_income` will be removed from the model.

Additionally, other income related factors that are generally associated with integration were examined (see **Appendix B.3**). The `Middle_class` variable is colinear with `Gini` and `Gini_99`, while the `Share01` variable is co-linear with `Gini`. Additionally, `Gini` seems to be highly predictive of `Gini_99`. `Income` is not strongly associated with any of the other variables examined.

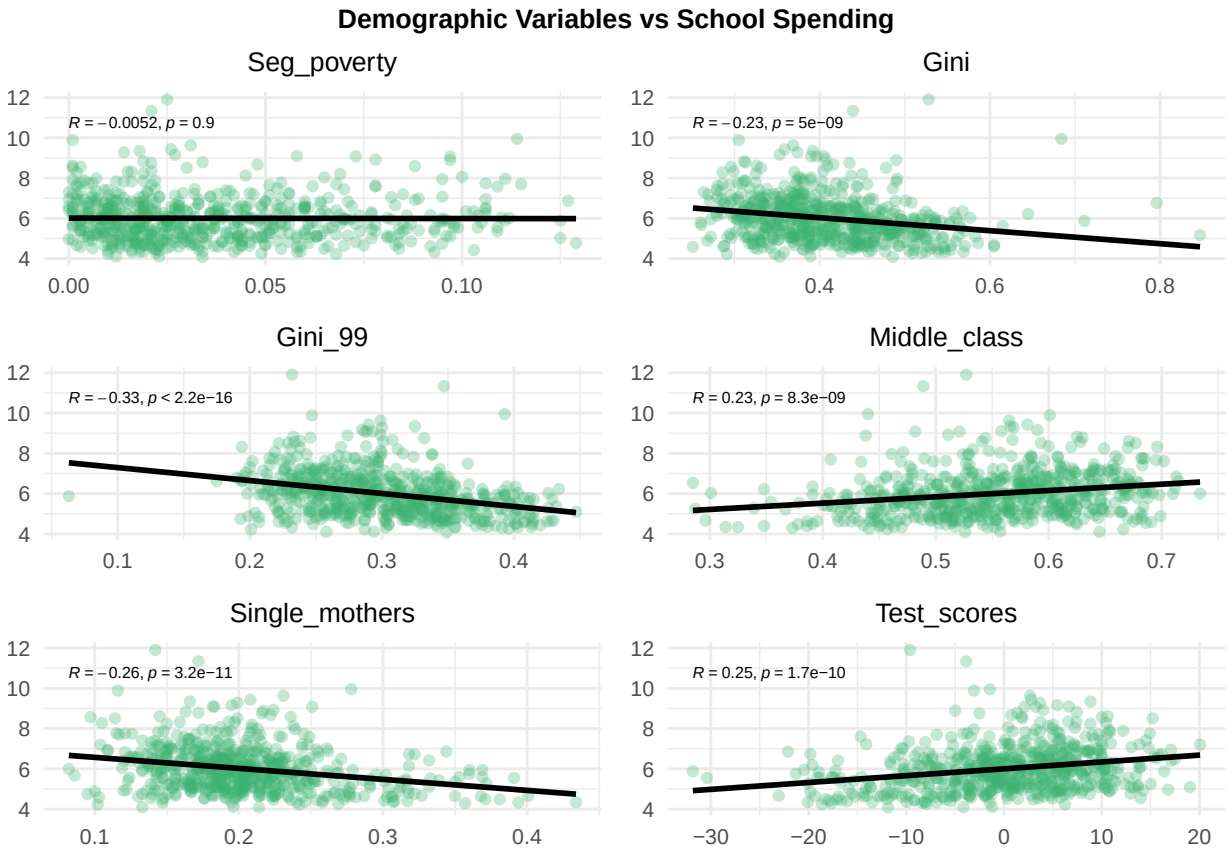
Segregation as a predictor of mobility

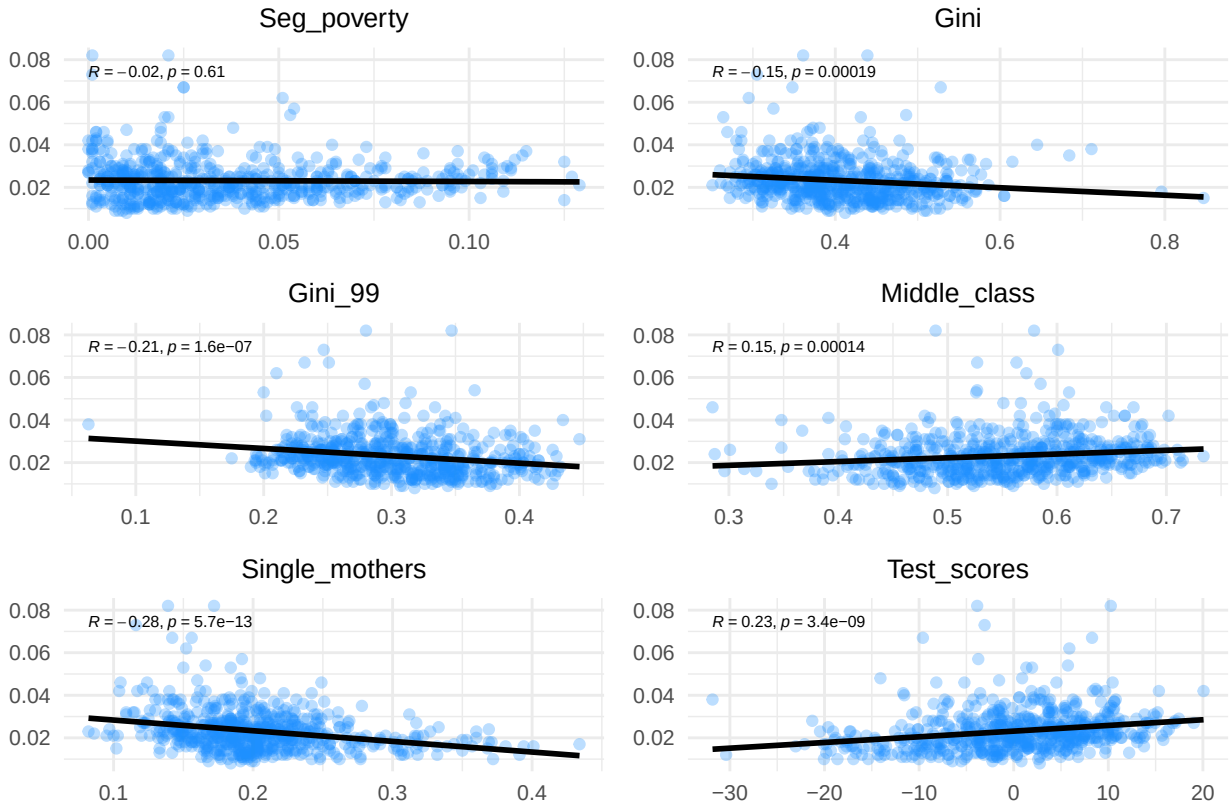


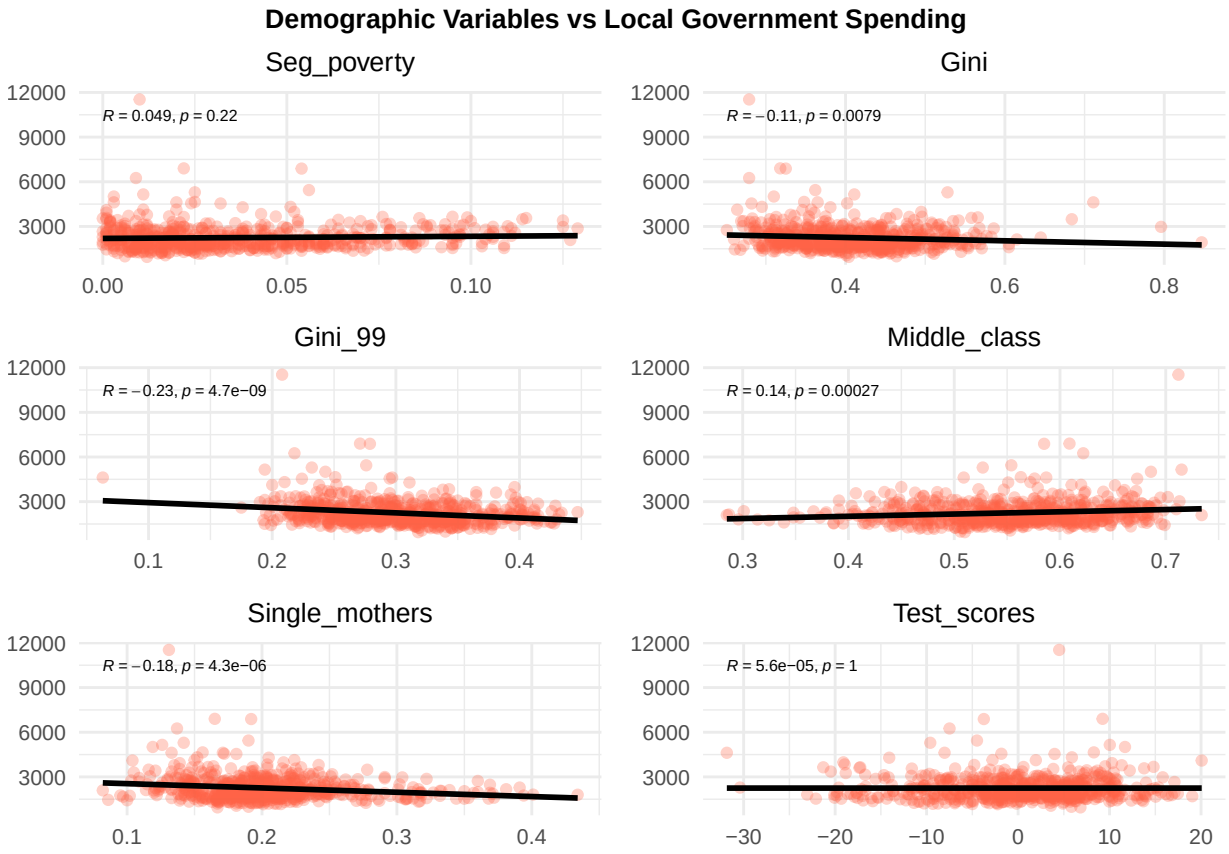
Appendices

Appendix A

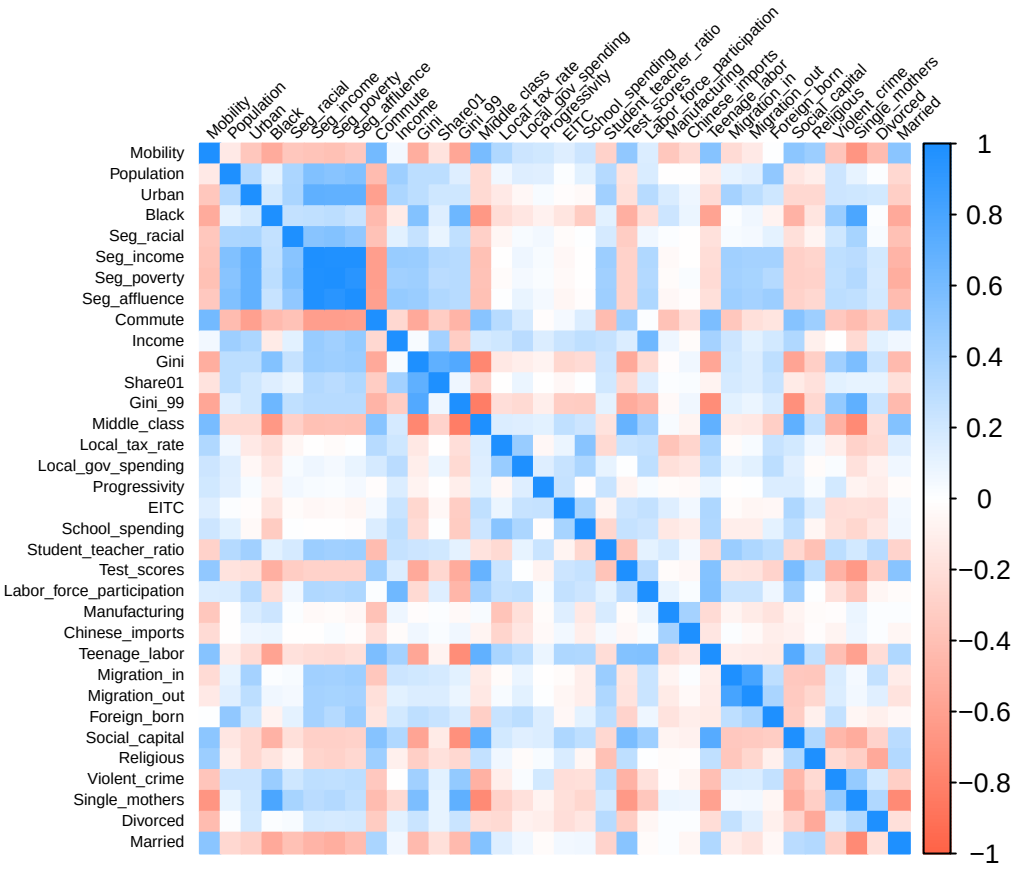
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```


Demographic Variables vs School Spending
Seg_poverty
R = −0.0052, p = 0.9
Gini
R = −0.23, p = 5e−09
Gini_99
R = −0.33, p < 2.2e−16
Middle_class
R = 0.23, p = 8.3e−09
Single_mothers
R = −0.26, p = 3.2e−11
Test_scores
R = 0.25, p = 1.7e−10

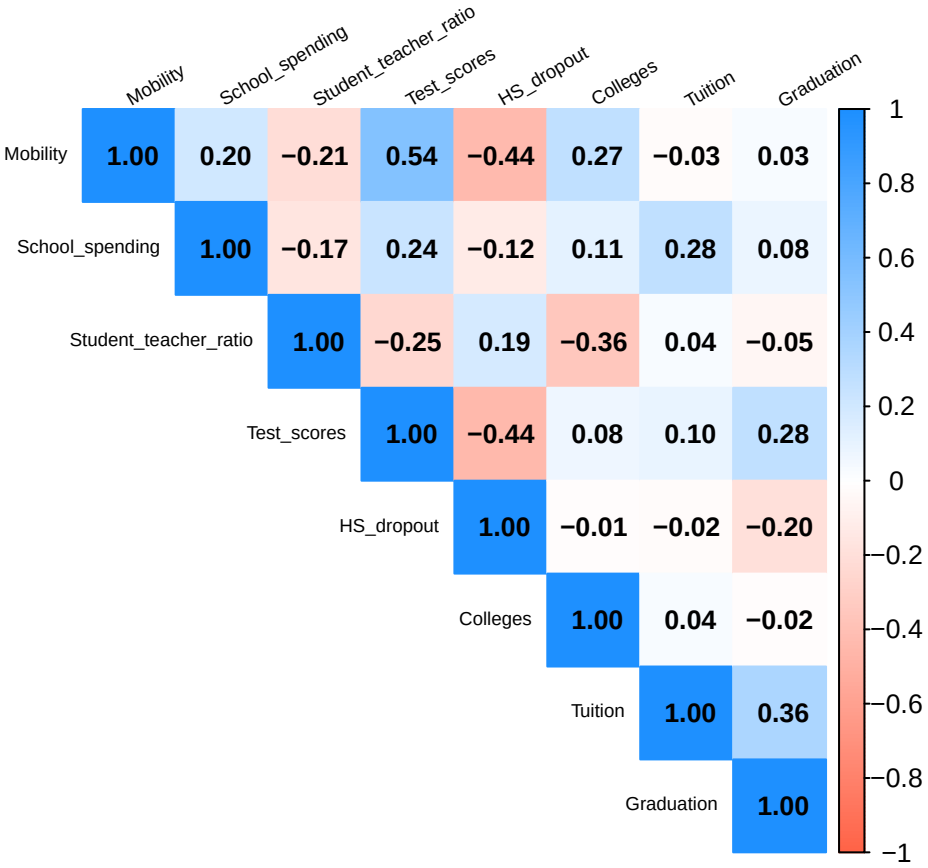

Demographic Variables vs Local Tax Rate
Seg_poverty
R = −0.02, p = 0.61
Gini
R = −0.15, p = 0.00019
Gini_99
R = −0.21, p = 1.6e−07
Middle_class
R = 0.15, p = 0.00014
Single_mothers
R = −0.28, p = 5.7e−13
Test_scores
R = 0.23, p = 3.4e−09


Demographic Variables vs Local Government Spending
Seg_poverty
R = 0.049, p = 0.22
Gini
R = −0.11, p = 0.0079
Gini_99
R = −0.23, p = 4.7e−09
Middle_class
R = 0.14, p = 0.00027
Single_mothers
R = −0.18, p = 4.3e−06
Test_scores
R = 5.6e−05, p = 1

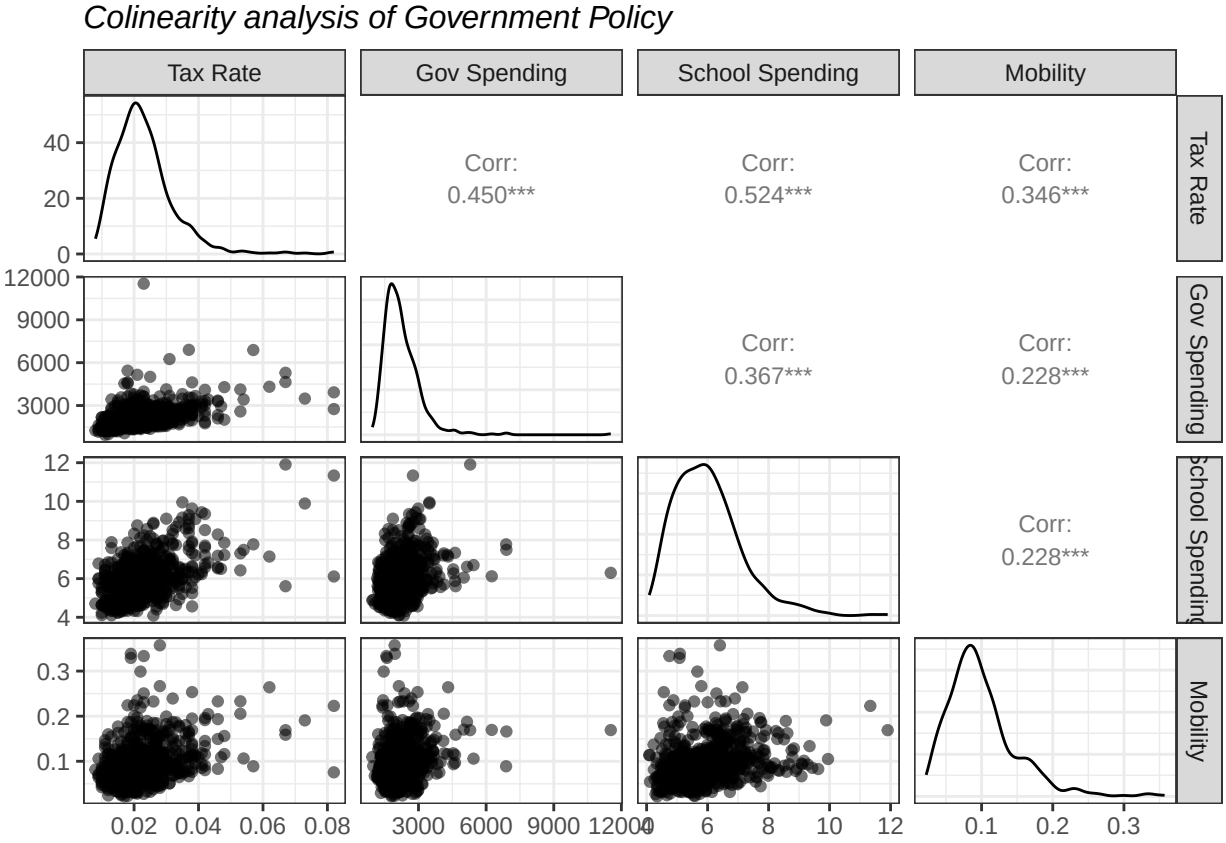
Appendix B - Colinearity Analysis



1 Education variables

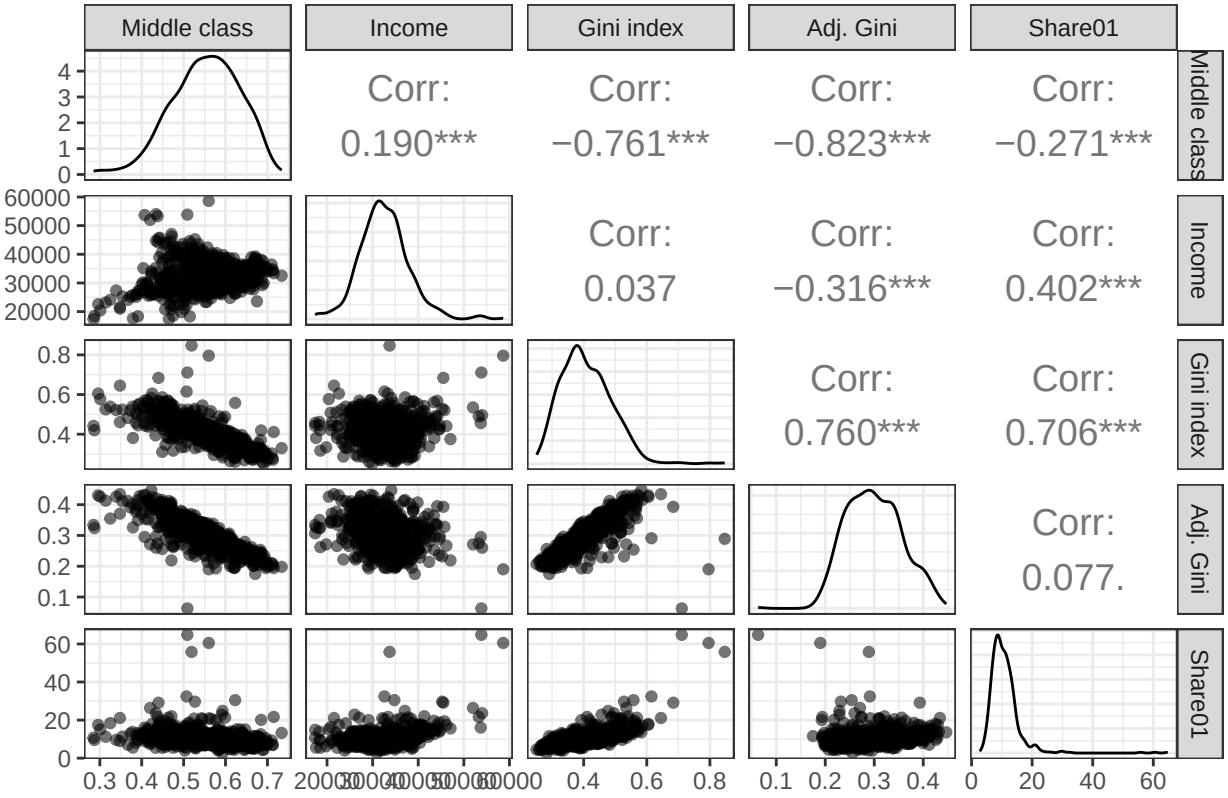


2 Government policy variables



3 Inequality variables

Colinearity analysis of income and income inequality



4 Segregation variables