Products of Our Environment

Amie Thomas

Abstract

The objective of the project was to investigate the relationship between social infrastructure (e.g., libraries, casinos) within neighborhoods and the likelihood of residents in those neighborhoods acquiring a college degree. This study utilized data from four datasets obtained from the National Neighborhood Data Archive, which were preprocessed, merged, and cleaned to form a comprehensive dataset. Following data preparation, exploratory data analysis was conducted to discern the most appropriate modeling approach. A hierarchical logistic regression model was selected due to its suitability in handling hierarchical data structures and binary outcomes. Multiple models were developed and compared based on their accuracy, precision, and F1-score to determine the most effective model in predicting college degree acquisition concerning neighborhood social infrastructure. The chosen model aimed to provide insights into the nuanced relationship between social infrastructure and educational attainment within neighborhoods, offering a comprehensive understanding of how these factors interrelate.

Introduction

This project explores how neighborhood amenities, like libraries and entertainment centers, influence residents' educational outcomes. Inspired by theories like Bronfenbrenner's Ecological Systems Theory and Shaw and McKay's Social Disorganization Theory, it delves into how neighborhoods shape individuals' lives. We predict that areas with more vice businesses might correlate with lower education levels, while those with learning and entertainment centers could support higher educational achievements. By analyzing neighborhood attributes and educational outcomes, we aim to understand how communities affect access to education and opportunities, aligning with theories on human development within community contexts.

Methods

I collected data from the National Neighborhood and Data Archive (NanDA), a repository offering various measures of the physical, economic, demographic, and social environment across

different spatial scales. The datasets encompassed socio-economic status, demographic information, neighborhood amenities, as well as specific data on Liquor, Tobacco, and Convenience Stores, and social and religious organizations. Initially, the datasets were filtered for the year 2017, which was the focal point of interest due to its relevance and timeliness.

To ensure a fair representation of neighborhoods, I retained only columns indicating social infrastructure per 1000 residents. This adjustment aimed to normalize smaller neighborhood sizes for equitable analysis alongside larger neighborhoods. Subsequently, I conducted a check for missing values within the dataset. Notably, several rows exhibited either 0's or NA's for population counts within a census tract. Given their lack of informative value for analysis, these rows were removed from consideration.

Furthermore, I associated each census tract with its respective census region. Finally, all datasets were consolidated, utilizing the census tract FIPS code as a common identifier for combining information across multiple datasets.

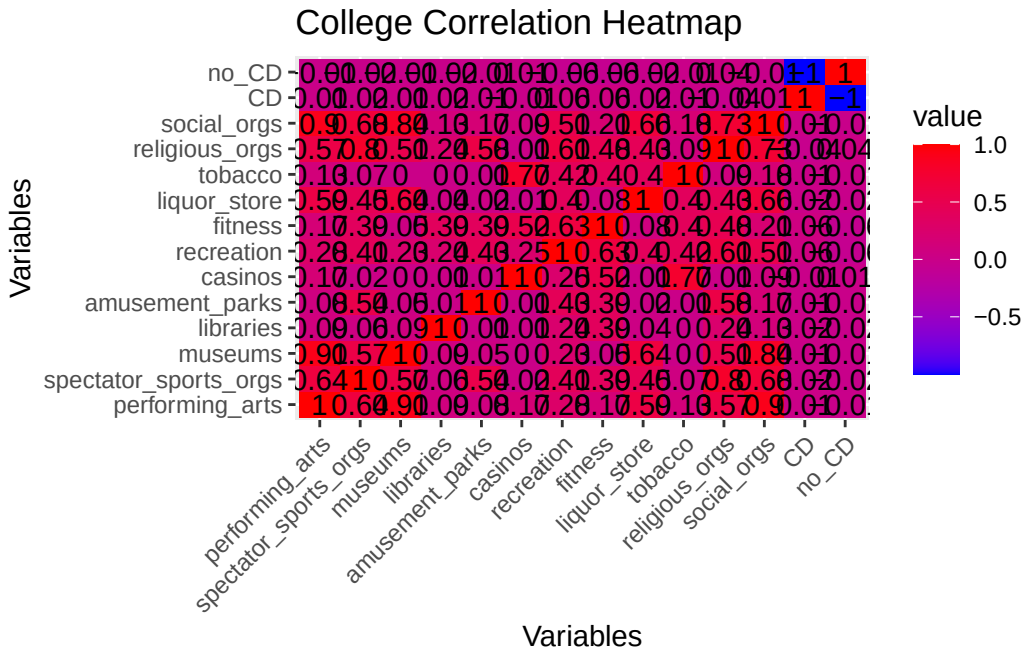
Following data collection, an exploratory data analysis was conducted to ascertain the most suitable model for the analysis. Presented below are the summarized outcomes derived from the dataset exploration.

Table 1: Summary of Data

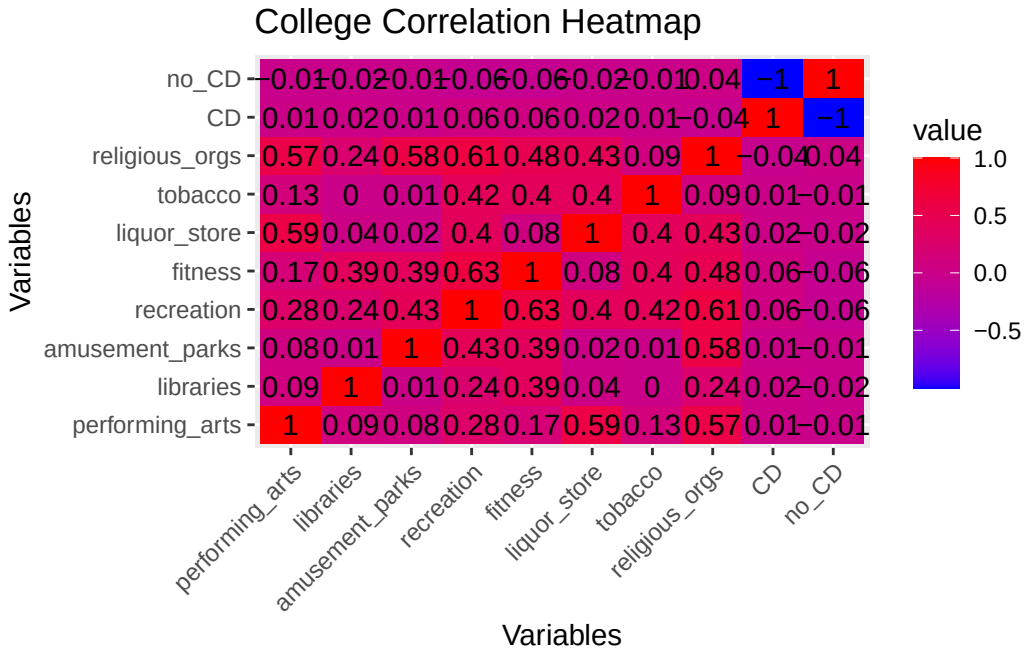
tract	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	college_edu
Length:72410	Length:72410	Length:72410	Min. :2017	Min. : 3	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.0000	Min. : 0.000	Min. :0.0000	Min. :0.0000	Min. :0.00000
Class :character	Class :character	Class :character	1st Qu.:2017	1st Qu.: 2946	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.2276	1st Qu.: 6.0000	1st Qu.: 0.0000	1st Qu.: 0.0000	1st Qu.: 0.6204	1st Qu.: 4.000	1st Qu.:0.1490	1st Qu.:0.3593	1st Qu.:0.30000
Mode :character	Mode :character	Mode :character	Median :2017	Median : 4131	Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.0000	Median : 0.5444	Median : 1.1589	Median : 0.0000	Median : 0.0000	Median : 1.1702	Median : 2.353	Median :0.2450	Median :0.5775	Median :0.40000
NA	NA	NA	Mean :2017	Mean : 4433	Mean : 0.2472	Mean : 2.0884	Mean : 6.1744	Mean : 4.1288	Mean : 8.0218	Mean : 8.8156	Mean : 6.8628	Mean : 8.3222	Mean : 2.1532	Mean : 2.2054	Mean : 1.3577	Mean : 6.715	Mean :0.2962	Mean :0.6270	Mean :0.81597
NA	NA	NA	3rd Qu.:2017	3rd Qu.: 5531	3rd Qu.: 0.2791	3rd Qu.: 1.0000	3rd Qu.: 0.1325	3rd Qu.: 0.1614	3rd Qu.: 4.0000	3rd Qu.: 0.0001	3rd Qu.: 1.0002	3rd Qu.: 3.4134	3rd Qu.: 0.2184	3rd Qu.: 4.0001	3rd Qu.: 0.9673	3rd Qu.: 3.712	3rd Qu.:0.4067	3rd Qu.:0.8507	3rd Qu.:0.70000
NA	NA	NA	Max. :2017	Max. :65528	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :[illegible]	Max. :1.0000	Max. :1.0000	Max. :1.00000

Following the data summary, I conducted a correlation analysis to explore potential correlations exceeding 0.70 among the predictor variables. The correlation matrix revealed notable correlations between variables such as social organization, casinos, museums, and spectator sports. These specific variables underwent comparison, and the one demonstrating the weakest correlation with the response variable was subsequently removed. This step aimed to mitigate the potential presence of multicollinearity in the analysis, ensuring a more robust model assessment.

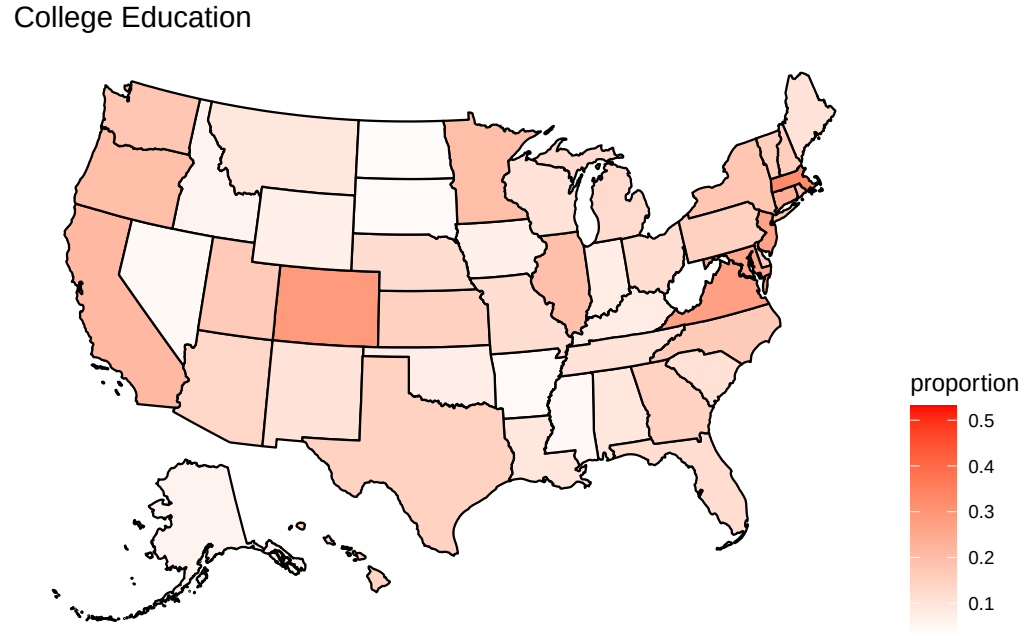
Before Variable Removal



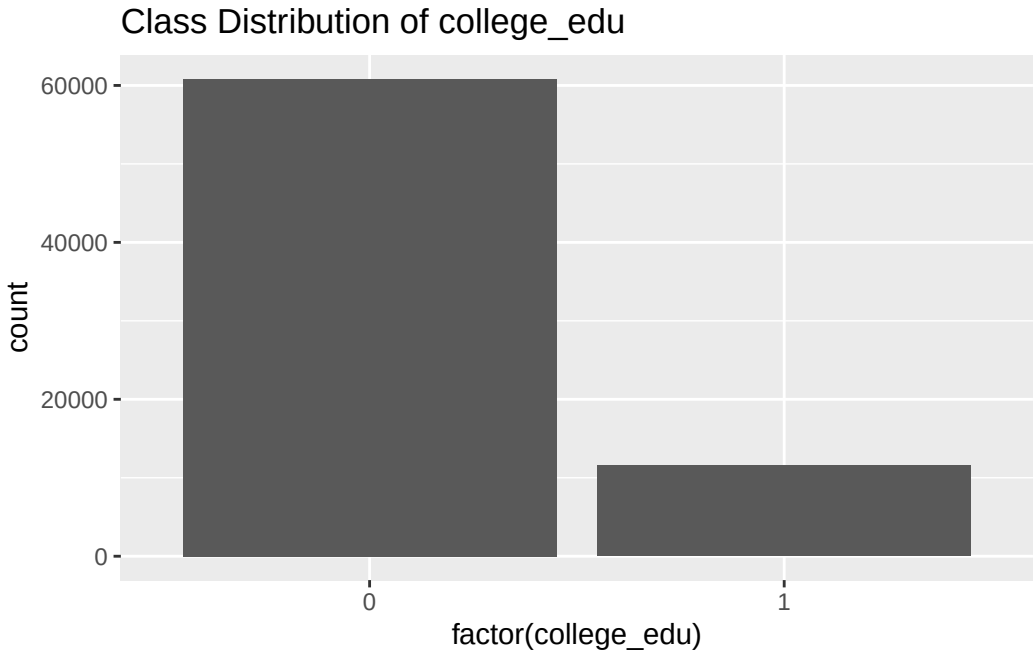
After Variable Removal



A visual representation of the United States was generated, depicting the proportion of residents with a college education by state. Noticeable variations in coloration among states were observed, indicating distinct differences between geographical groups. These discrepancies in educational attainment across states emphasize the necessity for employing a hierarchical model, considering the apparent dissimilarities and potential group-level effects within the dataset.



Finally, an evaluation of the response variable for class imbalance was conducted. To address this imbalance, a combination of undersampling and oversampling methods was employed, ensuring a more balanced representation within the dataset.



```
    0     1
36224 36186


Call:
glm(formula = college_edu ~ 1, family = binomial, data = train_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.178  -1.178   1.177   1.177   1.177

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0009667  0.0083097   0.116    0.907

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80305  on 57927  degrees of freedom
Residual deviance: 80305  on 57927  degrees of freedom
AIC: 80307

Number of Fisher Scoring iterations: 2

[1] "Misclassification Error: 0.502278690788565"

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred


Call:
glm(formula = college_edu ~ performing_arts + libraries + amusement_parks +
    recreation + fitness + liquor_store + tobacco + religious_orgs,
    family = binomial(link = "logit"), data = train_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.490  -1.010   0.000   1.059   8.490

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.308506   0.016279 -18.952  < 2e-16 ***
```

```
performing_arts  1.217368   0.032194  37.813  < 2e-16 ***
libraries        0.240761   0.036064   6.676 2.46e-11 ***
amusement_parks -0.290154   0.066875  -4.339 1.43e-05 ***
recreation       0.315361   0.015501  20.344  < 2e-16 ***
fitness          0.990728   0.031077  31.880  < 2e-16 ***
liquor_store    -0.021420   0.037824  -0.566    0.571
tobacco         -0.573864   0.072833  -7.879 3.29e-15 ***
religious_orgs  -0.460428   0.009614 -47.892  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80305  on 57927  degrees of freedom
Residual deviance: 69982  on 57919  degrees of freedom
AIC: 70000

Number of Fisher Scoring iterations: 8


[1] "Misclassification Error: 0.29899185195415"


[1] 0.7165859

Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred



Call:
glm(formula = college_edu ~ libraries + amusement_parks + liquor_store +
    tobacco + factor(state), family = binomial(link = "logit"),
    data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2980  -1.1201   0.0136   1.1128   2.1119

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.59050    0.07182  -8.222  < 2e-16 ***
libraries                  0.21568    0.02858   7.547 4.45e-14 ***
amusement_parks            0.37781    0.09041   4.179 2.93e-05 ***
liquor_store               0.24392    0.03376   7.224 5.04e-13 ***
```

```
tobacco                    -0.06166    0.05118  -1.205 0.228298
factor(state)Alaska        -0.73512    0.24355  -3.018 0.002541 **
factor(state)Arizona        0.40374    0.09268   4.356 1.32e-05 ***
factor(state)Arkansas      -1.02073    0.15620  -6.535 6.38e-11 ***
factor(state)California     0.89058    0.07571  11.763  < 2e-16 ***
factor(state)Colorado       1.33667    0.09428  14.178  < 2e-16 ***
factor(state)Connecticut    0.90225    0.10283   8.775  < 2e-16 ***
factor(state)DC             2.22974    0.19014  11.727  < 2e-16 ***
factor(state)Delaware       0.39633    0.17799   2.227 0.025970 *
factor(state)Florida        0.25015    0.08082   3.095 0.001966 **
factor(state)Georgia        0.45430    0.08841   5.138 2.77e-07 ***
factor(state)Hawaii         0.40265    0.14307   2.814 0.004888 **
factor(state)Idaho         -0.60682    0.19182  -3.164 0.001559 **
factor(state)Illinois       0.76772    0.08182   9.383  < 2e-16 ***
factor(state)Indiana       -0.26775    0.10147  -2.639 0.008320 **
factor(state)Iowa          -0.48994    0.12429  -3.942 8.08e-05 ***
factor(state)Kansas         0.37624    0.11071   3.398 0.000678 ***
factor(state)Kentucky      -0.15837    0.10876  -1.456 0.145364
factor(state)Louisiana     -0.16479    0.10594  -1.555 0.119847
factor(state)Maine          0.06748    0.15192   0.444 0.656902
factor(state)Maryland       1.08605    0.09149  11.871  < 2e-16 ***
factor(state)Massachusetts  1.48238    0.09045  16.389  < 2e-16 ***
factor(state)Michigan       0.20033    0.08516   2.352 0.018648 *
factor(state)Minnesota      0.80698    0.09251   8.723  < 2e-16 ***
factor(state)Mississippi   -0.84958    0.14907  -5.699 1.20e-08 ***
factor(state)Missouri       0.13555    0.09686   1.399 0.161672
factor(state)Montana       -0.29382    0.17757  -1.655 0.097978 .
factor(state)Nebraska       0.17434    0.12564   1.388 0.165242
factor(state)Nevada        -0.59255    0.13764  -4.305 1.67e-05 ***
factor(state)New Hampshire  0.19183    0.15070   1.273 0.203048
factor(state)New Jersey     1.13212    0.08490  13.334  < 2e-16 ***
factor(state)New Mexico    -0.02026    0.12983  -0.156 0.875977
factor(state)New York       0.63170    0.07856   8.041 8.93e-16 ***
factor(state)North Carolina 0.61699    0.08654   7.129 1.01e-12 ***
factor(state)North Dakota  -1.10787    0.26849  -4.126 3.69e-05 ***
factor(state)Ohio           0.15515    0.08468   1.832 0.066942 .
factor(state)Oklahoma      -0.32951    0.10904  -3.022 0.002512 **
factor(state)Oregon         0.92091    0.10333   8.912  < 2e-16 ***
factor(state)Pennsylvania   0.40978    0.08204   4.995 5.89e-07 ***
factor(state)Rhode Island   0.44054    0.15676   2.810 0.004950 **
factor(state)South Carolina 0.22060    0.10334   2.135 0.032781 *
factor(state)South Dakota  -1.51499    0.27566  -5.496 3.89e-08 ***
factor(state)Tennessee     -0.01959    0.09646  -0.203 0.839055
```

```
factor(state)Texas           0.42097    0.07836   5.372 7.79e-08 ***
factor(state)Utah            0.62049    0.11612   5.343 9.12e-08 ***
factor(state)Vermont         0.65370    0.17813   3.670 0.000243 ***
factor(state)Virginia        1.20962    0.08632  14.014  < 2e-16 ***
factor(state)Washington      0.63971    0.09219   6.939 3.95e-12 ***
factor(state)West Virginia  -1.52571    0.20672  -7.380 1.58e-13 ***
factor(state)Wisconsin      -0.08348    0.09880  -0.845 0.398117
factor(state)Wyoming        -0.65155    0.28812  -2.261 0.023737 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80305  on 57927  degrees of freedom
Residual deviance: 76560  on 57873  degrees of freedom
AIC: 76670

Number of Fisher Scoring iterations: 6


[1] "Misclassification Error: 0.402775859687888"


[1] 0.5896588


Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: college_edu ~ libraries + amusement_parks + liquor_store + tobacco +
    (1 | state)
   Data: train_data

     AIC      BIC   logLik deviance df.resid
 76844.7  76898.5 -38416.3  76832.7    57922

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1086.93    -0.93     0.01     0.93     2.72

Random effects:
 Groups Name        Variance Std.Dev.
 state  (Intercept) 0.516    0.7183
Number of obs: 57928, groups:  state, 51
```

```
Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.37979    0.10186  -3.728 0.000193 ***
libraries        0.21485    0.02853   7.530 5.07e-14 ***
amusement_parks  0.37644    0.09006   4.180 2.92e-05 ***
liquor_store     0.24597    0.03372   7.294 3.00e-13 ***
tobacco         -0.06244    0.05134  -1.216 0.223965
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "Misclassification Error: 0.40298301339594"

[1] 0.5895921

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: college_edu ~ libraries + amusement_parks + liquor_store + tobacco +
    (1 + libraries | state) + (1 + amusement_parks | state) +
    (1 + liquor_store | state) + (1 + tobacco | state)
   Data: train_data

     AIC      BIC   logLik deviance df.resid
 76167.7  76320.1 -38066.8  76133.7    57911

Scaled residuals:
     Min       1Q   Median       3Q      Max
-17.7054  -0.9102   0.0000   0.9268   6.9894

Random effects:
 Groups  Name            Variance Std.Dev. Corr
 state   (Intercept)     0.14306  0.3782
         libraries       1.57122  1.2535   0.72
 state.1 (Intercept)     0.15241  0.3904
         amusement_parks 2.45646  1.5673   -0.42
 state.2 (Intercept)     0.11100  0.3332
         liquor_store    0.54371  0.7374   -0.56
 state.3 (Intercept)     0.06852  0.2618
         tobacco         1.25486  1.1202   0.69
Number of obs: 57928, groups:  state, 51

Fixed effects:
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.33141    0.09759  -3.396 0.000684 ***
libraries       -0.03599    0.18471  -0.195 0.845532
amusement_parks  0.62046    0.26717   2.322 0.020215 *
liquor_store     0.25405    0.12297   2.066 0.038828 *
tobacco         -0.45228    0.19664  -2.300 0.021445 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "Misclassification Error: 0.415412235879022"

[1] 0.5545683

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: college_edu ~ libraries + amusement_parks + liquor_store + tobacco +
    (libraries | state) + (amusement_parks | state) + (liquor_store |
    state) + (tobacco | state)
   Data: train_data

     AIC      BIC   logLik deviance df.resid
 76167.7  76320.1 -38066.8  76133.7    57911

Scaled residuals:
     Min       1Q   Median       3Q      Max
-17.7054  -0.9102   0.0000   0.9268   6.9894

Random effects:
 Groups  Name            Variance Std.Dev. Corr
 state   (Intercept)     0.14306  0.3782
         libraries       1.57122  1.2535   0.72
 state.1 (Intercept)     0.15241  0.3904
         amusement_parks 2.45646  1.5673   -0.42
 state.2 (Intercept)     0.11100  0.3332
         liquor_store    0.54371  0.7374   -0.56
 state.3 (Intercept)     0.06852  0.2618
         tobacco         1.25486  1.1202   0.69
Number of obs: 57928, groups:  state, 51

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)     -0.33141    0.09759  -3.396 0.000684 ***
libraries       -0.03599    0.18471  -0.195 0.845532
amusement_parks  0.62046    0.26717   2.322 0.020215 *
liquor_store     0.25405    0.12297   2.066 0.038828 *
tobacco         -0.45228    0.19664  -2.300 0.021445 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
[1] "Misclassification Error: 0.415412235879022"
```

```
[1] 0.5545683
```

Table: AIC values of Models

Model	AIC
Null model	80307.25
Complete pooling model	69999.87
No pooling model	76669.83
Partial pooling model vary intercept	76844.67
Partial pooling model vary both	76167.70
Partial pooling model vary slope	76167.70

Table: Misclassification Errors of Models

Model	Misclassification.Error
Null model	0.50
Complete pooling model	0.30
No pooling model	0.40
Partial pooling model vary intercept	0.40
Partial pooling model vary both	0.42
Partial pooling model vary slope	0.41

Table: F1 test scores of Models

Model	F1.Test.Scores
Complete pooling model	0.72
No pooling model	0.59
Partial pooling model vary intercept	0.59
Partial pooling model vary both	0.55
Partial pooling model vary slope	0.55

Following the preparatory steps, the subsequent phase involved constructing the models.

The Null Model

```
Call:
glm(formula = college_edu ~ 1, family = binomial, data = train_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.178  -1.178   1.177   1.177   1.177

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.0009667  0.0083097   0.116    0.907

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80305  on 57927  degrees of freedom
Residual deviance: 80305  on 57927  degrees of freedom
AIC: 80307

Number of Fisher Scoring iterations: 2
```

Complete Pooling Model

```
Call:
glm(formula = college_edu ~ performing_arts + libraries + amusement_parks +
```

```
    recreation + fitness + liquor_store + tobacco + religious_orgs,
    family = binomial(link = "logit"), data = train_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.490  -1.010   0.000   1.059   8.490

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.308506   0.016279 -18.952  < 2e-16 ***
performing_arts  1.217368   0.032194  37.813  < 2e-16 ***
libraries        0.240761   0.036064   6.676 2.46e-11 ***
amusement_parks -0.290154   0.066875  -4.339 1.43e-05 ***
recreation       0.315361   0.015501  20.344  < 2e-16 ***
fitness          0.990728   0.031077  31.880  < 2e-16 ***
liquor_store    -0.021420   0.037824  -0.566    0.571
tobacco         -0.573864   0.072833  -7.879 3.29e-15 ***
religious_orgs  -0.460428   0.009614 -47.892  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80305  on 57927  degrees of freedom
Residual deviance: 69982  on 57919  degrees of freedom
AIC: 70000

Number of Fisher Scoring iterations: 8
```

No Pooling Model

```
Call:
glm(formula = college_edu ~ libraries + amusement_parks + liquor_store +
    tobacco + factor(state), family = binomial(link = "logit"),
    data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2980  -1.1201   0.0136   1.1128   2.1119

Coefficients:
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.59050	0.07182	-8.222	< 2e-16	***
libraries	0.21568	0.02858	7.547	4.45e-14	***
amusement_parks	0.37781	0.09041	4.179	2.93e-05	***
liquor_store	0.24392	0.03376	7.224	5.04e-13	***
tobacco	-0.06166	0.05118	-1.205	0.228298	
factor(state)Alaska	-0.73512	0.24355	-3.018	0.002541	**
factor(state)Arizona	0.40374	0.09268	4.356	1.32e-05	***
factor(state)Arkansas	-1.02073	0.15620	-6.535	6.38e-11	***
factor(state)California	0.89058	0.07571	11.763	< 2e-16	***
factor(state)Colorado	1.33667	0.09428	14.178	< 2e-16	***
factor(state)Connecticut	0.90225	0.10283	8.775	< 2e-16	***
factor(state)DC	2.22974	0.19014	11.727	< 2e-16	***
factor(state)Delaware	0.39633	0.17799	2.227	0.025970	*
factor(state)Florida	0.25015	0.08082	3.095	0.001966	**
factor(state)Georgia	0.45430	0.08841	5.138	2.77e-07	***
factor(state)Hawaii	0.40265	0.14307	2.814	0.004888	**
factor(state)Idaho	-0.60682	0.19182	-3.164	0.001559	**
factor(state)Illinois	0.76772	0.08182	9.383	< 2e-16	***
factor(state)Indiana	-0.26775	0.10147	-2.639	0.008320	**
factor(state)Iowa	-0.48994	0.12429	-3.942	8.08e-05	***
factor(state)Kansas	0.37624	0.11071	3.398	0.000678	***
factor(state)Kentucky	-0.15837	0.10876	-1.456	0.145364	
factor(state)Louisiana	-0.16479	0.10594	-1.555	0.119847	
factor(state)Maine	0.06748	0.15192	0.444	0.656902	
factor(state)Maryland	1.08605	0.09149	11.871	< 2e-16	***
factor(state)Massachusetts	1.48238	0.09045	16.389	< 2e-16	***
factor(state)Michigan	0.20033	0.08516	2.352	0.018648	*
factor(state)Minnesota	0.80698	0.09251	8.723	< 2e-16	***
factor(state)Mississippi	-0.84958	0.14907	-5.699	1.20e-08	***
factor(state)Missouri	0.13555	0.09686	1.399	0.161672	
factor(state)Montana	-0.29382	0.17757	-1.655	0.097978	.
factor(state)Nebraska	0.17434	0.12564	1.388	0.165242	
factor(state)Nevada	-0.59255	0.13764	-4.305	1.67e-05	***
factor(state)New Hampshire	0.19183	0.15070	1.273	0.203048	
factor(state)New Jersey	1.13212	0.08490	13.334	< 2e-16	***
factor(state)New Mexico	-0.02026	0.12983	-0.156	0.875977	
factor(state)New York	0.63170	0.07856	8.041	8.93e-16	***
factor(state)North Carolina	0.61699	0.08654	7.129	1.01e-12	***
factor(state)North Dakota	-1.10787	0.26849	-4.126	3.69e-05	***
factor(state)Ohio	0.15515	0.08468	1.832	0.066942	.
factor(state)Oklahoma	-0.32951	0.10904	-3.022	0.002512	**
factor(state)Oregon	0.92091	0.10333	8.912	< 2e-16	***

```
factor(state)Pennsylvania    0.40978    0.08204   4.995 5.89e-07 ***
factor(state)Rhode Island    0.44054    0.15676   2.810 0.004950 **
factor(state)South Carolina  0.22060    0.10334   2.135 0.032781 *
factor(state)South Dakota   -1.51499    0.27566  -5.496 3.89e-08 ***
factor(state)Tennessee      -0.01959    0.09646  -0.203 0.839055
factor(state)Texas           0.42097    0.07836   5.372 7.79e-08 ***
factor(state)Utah            0.62049    0.11612   5.343 9.12e-08 ***
factor(state)Vermont         0.65370    0.17813   3.670 0.000243 ***
factor(state)Virginia        1.20962    0.08632  14.014  < 2e-16 ***
factor(state)Washington      0.63971    0.09219   6.939 3.95e-12 ***
factor(state)West Virginia  -1.52571    0.20672  -7.380 1.58e-13 ***
factor(state)Wisconsin      -0.08348    0.09880  -0.845 0.398117
factor(state)Wyoming        -0.65155    0.28812  -2.261 0.023737 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80305  on 57927  degrees of freedom
Residual deviance: 76560  on 57873  degrees of freedom
AIC: 76670

Number of Fisher Scoring iterations: 6
```

Partial Pooling Model with Varying Intercept

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: college_edu ~ libraries + amusement_parks + liquor_store + tobacco +
    (1 | state)
   Data: train_data

     AIC      BIC   logLik deviance df.resid
 76844.7  76898.5 -38416.3  76832.7    57922

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1086.93    -0.93     0.01     0.93     2.72

Random effects:
 Groups Name        Variance Std.Dev.
```

```
 state  (Intercept) 0.516    0.7183
Number of obs: 57928, groups:  state, 51

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.37979    0.10186  -3.728 0.000193 ***
libraries        0.21485    0.02853   7.530 5.07e-14 ***
amusement_parks  0.37644    0.09006   4.180 2.92e-05 ***
liquor_store     0.24597    0.03372   7.294 3.00e-13 ***
tobacco         -0.06244    0.05134  -1.216 0.223965
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial Pooling Model with Varying Slope

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: college_edu ~ libraries + amusement_parks + liquor_store + tobacco +
    (libraries | state) + (amusement_parks | state) + (liquor_store |
    state) + (tobacco | state)
   Data: train_data

     AIC      BIC   logLik deviance df.resid
 76167.7  76320.1 -38066.8  76133.7    57911

Scaled residuals:
     Min       1Q   Median       3Q      Max
-17.7054  -0.9102   0.0000   0.9268   6.9894

Random effects:
 Groups  Name            Variance Std.Dev. Corr
 state   (Intercept)     0.14306  0.3782
         libraries       1.57122  1.2535   0.72
 state.1 (Intercept)     0.15241  0.3904
         amusement_parks 2.45646  1.5673   -0.42
 state.2 (Intercept)     0.11100  0.3332
         liquor_store    0.54371  0.7374   -0.56
 state.3 (Intercept)     0.06852  0.2618
         tobacco         1.25486  1.1202   0.69
Number of obs: 57928, groups:  state, 51
```

```
Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.33141    0.09759  -3.396 0.000684 ***
libraries       -0.03599    0.18471  -0.195 0.845532
amusement_parks  0.62046    0.26717   2.322 0.020215 *
liquor_store     0.25405    0.12297   2.066 0.038828 *
tobacco         -0.45228    0.19664  -2.300 0.021445 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Partial Pooling Model with Varying Intercept and Varying Slope

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: college_edu ~ libraries + amusement_parks + liquor_store + tobacco +
    (1 + libraries | state) + (1 + amusement_parks | state) +
    (1 + liquor_store | state) + (1 + tobacco | state)
   Data: train_data

     AIC      BIC   logLik deviance df.resid
 76167.7  76320.1 -38066.8  76133.7    57911

Scaled residuals:
     Min       1Q   Median       3Q      Max
-17.7054  -0.9102   0.0000   0.9268   6.9894

Random effects:
 Groups  Name            Variance Std.Dev. Corr
 state   (Intercept)     0.14306  0.3782
         libraries       1.57122  1.2535   0.72
 state.1 (Intercept)     0.15241  0.3904
         amusement_parks 2.45646  1.5673   -0.42
 state.2 (Intercept)     0.11100  0.3332
         liquor_store    0.54371  0.7374   -0.56
 state.3 (Intercept)     0.06852  0.2618
         tobacco         1.25486  1.1202   0.69
Number of obs: 57928, groups:  state, 51

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.33141    0.09759  -3.396 0.000684 ***
```

```
libraries       -0.03599    0.18471  -0.195 0.845532
amusement_parks  0.62046    0.26717   2.322 0.020215 *
liquor_store     0.25405    0.12297   2.066 0.038828 *
tobacco         -0.45228    0.19664  -2.300 0.021445 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The models were compared based on their misclassification errors, which measure how often the models predict the wrong class compared to all predictions made. A lower error means the models are more accurate, while a higher error suggests they make more incorrect predictions. This comparison helps identify areas where the models could be improved.

```
print(missclass_ktable)
```

Table: Misclassification Errors of Models

Model	Misclassification.Error
Null model	0.50
Complete pooling model	0.30
No pooling model	0.40
Partial pooling model vary intercept	0.40
Partial pooling model vary both	0.42
Partial pooling model vary slope	0.41

Next, the model's F1 scores were compared to give us a balanced view of how well the model works by considering both its precision and recall. They show how accurately the model identifies positive instances while reducing mistakes like false positives and false negatives. A higher F1 score means the model performs better overall.

```
print(f1_ktable)
```

Table: F1 test scores of Models

Model	F1.Test.Scores

Complete pooling model	0.72
No pooling model	0.59
Partial pooling model vary intercept	0.59
Partial pooling model vary both	0.55
Partial pooling model vary slope	0.55

Lastly the model's AIC values were compared to help us compare models by finding a sweet spot between accurately describing data and keeping things simple. Lower AIC values mean better models that strike a good balance between accuracy and simplicity.

```
print(aic_ktable)
```

Table: AIC values of Models

Model	AIC
Null model	80307.25
Complete pooling model	69999.87
No pooling model	76669.83
Partial pooling model vary intercept	76844.67
Partial pooling model vary both	76167.70
Partial pooling model vary slope	76167.70

According to the model validation and comparison tests run above, we can see that the complete pooling model performs better than all other models. We will interpret this model for our analysis.

Results

The results of the complete pooling model suggest performing arts, libraries, amusement parks, recreation, fitness, tobacco, and religious organisations appear statistically significant ($p < 0.05$), meaning they likely have a significant impact on the likelihood of the resident in the neighborhood have a college education. Further,

Performing arts: For a one-unit increase in performing arts, the log-odds of having a college education increase by 1.217.

Libraries: A one-unit increase in libraries is associated with an increase in the log-odds of the having a college education by 0.241.

Amusement parks: A one-unit increase in amusement_parks is linked to a decrease in the log-odds of having a college education 0.290.

Recreation: Each unit increase in recreation corresponds to an increase in the log-odds of having a college by 0.315.

Fitness: An increase in fitness by one unit is associated with a rise in the log-odds of having a college education by 0.991.

Liquor store: The variable liquor store does not significantly impact the log-odds of having a college education, given its non-significant p-value ($p = 0.571$).

Tobacco: For each unit increase in tobacco, the log-odds of having a college education decrease by 0.574.

Religious organizations: An increase in religious organizations by one unit leads to a decrease in the log-odds of having a college education by 0.460.

Discussion

The outcomes of this project present intriguing findings that contradict initial assumptions. Subsequent analyses will incorporate time series data to explore the impact of changes in social infrastructure on educational levels. This future analysis intends to introduce a lag to examine the delayed effects, aiming to deepen our understanding of the relationship between social infrastructure and education over time.

Appendix

Article used for guidance:

Author links open overlay panelTimothy Fraser a, a, c, d, b, AbstractScholars and policymakers increasingly recognize the value of social capital - the connections that generate and enable trust among people - in responding to and recovering from shocks and disasters. However, Altschuler, A., Brueckner, J. K., Fraser, T., Hanibuchi, T., Johnson, C. A., Krekel, C., Maas, J., O'Sullivan, T. L., Page-Tan, C., Skjaeveland, O., Aldrich, D. P., Aldrich, D. P., Alesina, A., … Follman, A. (2022, September 29). *Trust but verify: Validating new measures for mapping social infrastructure in cities*. Urban Climate. https://www.sciencedirect.com/science/article/abs/pii/S221209552200205X