LM Model Creation

Andrew Mendez

2025-02-10

Intro

We as a team were tasked with investigating economic mobility across decades within the united states. By looking at variables linked to social climate, economic prosperity, government policy, and education we were able to create a linear model but before we were able too create this model we had to perform essential exploratory data analysis. We were looking for variables that either with or without transformation met the assumptions necessary to build a linear model. This was a struggle for some variables due to their lack of information or more explicitly high amounts of NA values, these were mainly in the education based columns and none seemed to be a good fit for a linear model. We then used the variables we found best fit our linear model assumptions and used a stepwise method of variable selection comparing models based on their AIC. Lastly we compared how are model fared across regions and ultimately selected our best model. Using this model and what it deems to be good predictors of social mobility we hope to support government policy that will benefit our nation.

Overview of EDA

Bi-Variate analysis

Table 1 describes the Bi-Variate relationship between all numerical predictor variables to Mobility. This also include all relevant transformation to see if the predictor variable can represent a better linear relationship Mobility. The R^2 highlighted in red show the best value between all the transformations. To note NA values for transformations mean they were not applicable to the variable

Results

There are 4 values (Single Mothers, Commute, Gini, Black) have a R^2 of 0.30 or higher. These show Moderate to High correlations to Mobility.

Higher single motherhood rates correlate with lower mobility due to reduced household stability and fewer financial and educational resources for children. Commute times impact mobility because shorter commutes indicate better access to jobs, schools, and infrastructure, fostering economic opportunity. Income inequality (Gini coefficient) limits upward mobility as greater wealth gaps reduce access to quality education and social support systems. Lastly, the percentage of Black residents is associated with lower mobility due to systemic disparities, historical wealth gaps, and limited access to high-quality schools and job markets. These factors collectively shape economic opportunity and social mobility.

Multi-Variate analysis

Table 2 describes Multi-Variate relationship between all numerical predictor variables to Mobility. This also includes the VIF scores to check for colinearity. The Graphs are then for multiple ways of check for heteroscedasticity

Results

Looking alone at every variable there appears to be high colinearity between a lot of these variables, specifically around (Gini, Seg Affluence, Share 1%, Gini 99%, Seg Income, Seg Poverty). Since a lot of these variables don't follow the linear assumptions it would be better to cross a lot of these ones out in the final model.

Looking at the graphs it does show there is some heteroscedasticity happening here. Q-Q graph is not showing a constant straight line due to some outliers. Scale-Location appears also not to be a flat line either due to high residual high fitted values. This will be taken into account when model building.

One Theory due to high VIF scores on some variables is due to all of them sharing the common characteristic in that they relate to social groups. In that Gini is probably the best in explain the differences between social groups and all the other ones relate to Gini too.

NA Values EDA / Results

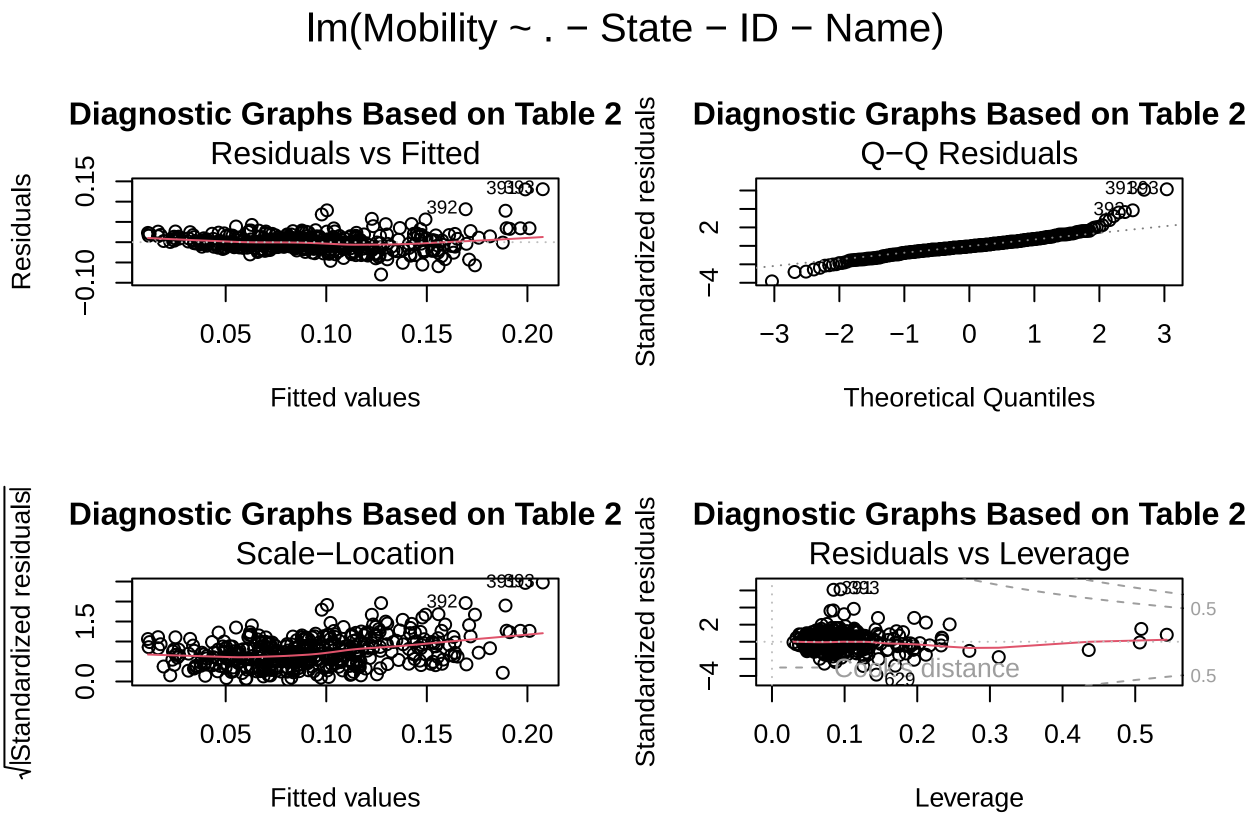
It appears that there is in fact a pattern in NA values. It appears that there's a lot more missing values in education related variables. Due to this and the high variance this will cause we will not be considering Education Varibales in the final model

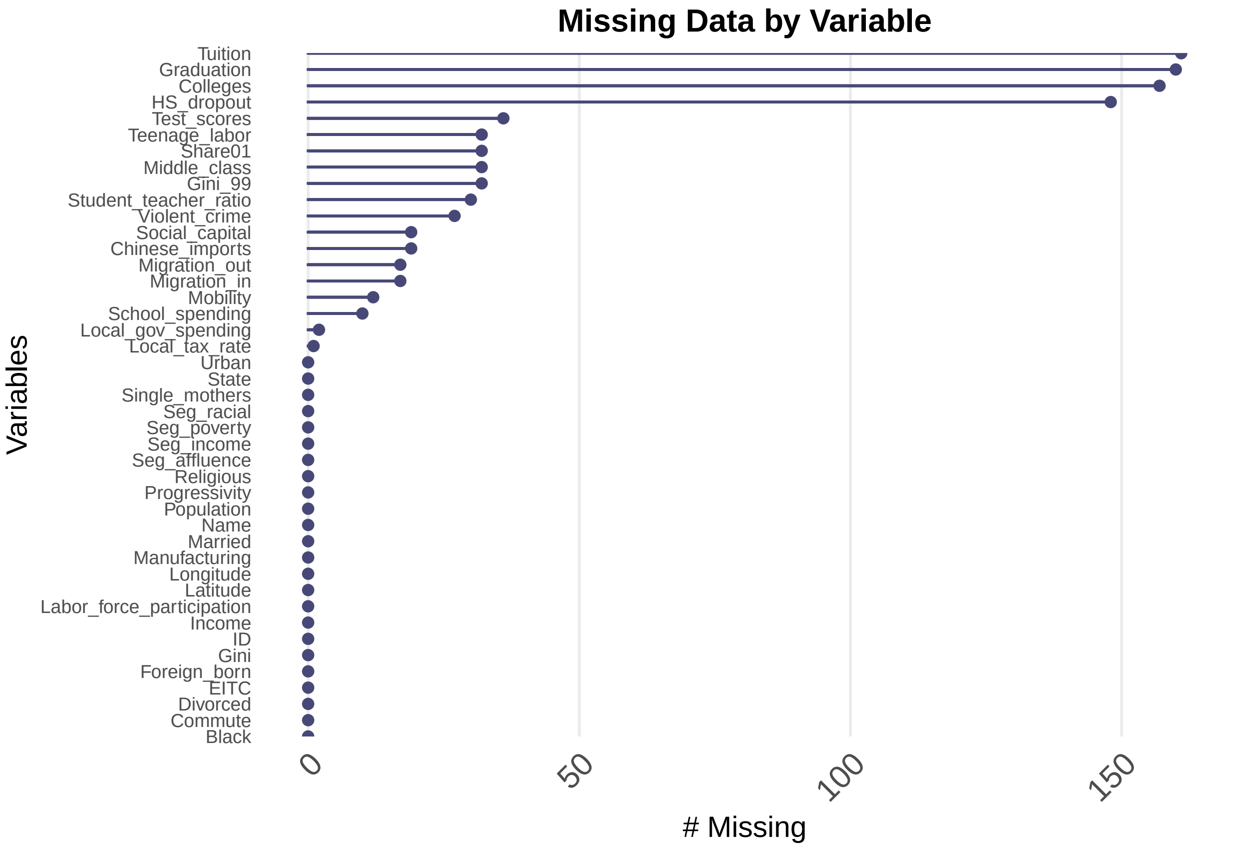
Table 1: Bivariate Importance Analysis

Predictor	Correlation	P-Val	R Squared	RSquared Log	RSquared Sqrt	RSquared Squared	RSquared Cube-root	RSquared Reciporcal
Single mothers	-0.6729	0.0000	0.4528	0.5116	0.4868	0.3696	0.4963	0.5236
Commute	0.5757	0.0000	0.3315	0.3089	0.3249	0.3168	0.3206	0.2518
Middle class	0.5212	0.0000	0.2717	0.251	0.2623	0.285	0.2588	0.2219
Gini 99	-0.5178	0.0000	0.2681	0.2556	0.2657	0.2619	0.2634	0.1849
Gini	-0.5159	0.0000	0.2662	0.2957	0.283	0.2214	0.2877	0.3078
Black	-0.5045	0.0000	0.2545	NA	0.3317	0.1617	0.3587	NA
Social capital	0.4984	0.0000	0.2484	NA	NA	0.0538	0.2074	0
Married	0.4970	0.0000	0.247	0.2347	0.2412	0.2567	0.2391	0.2196
Teenage labor	0.4821	0.0000	0.2324	0.2135	0.2252	0.2351	0.2218	0.1772
ID	0.4651	0.0000	0.2163	0.194	0.2255	0.1651	0.2214	0.0259
Religious	0.4246	0.0000	0.1802	0.1278	0.156	0.2109	0.1469	0.0679
Divorced	-0.4159	0.0000	0.173	0.2006	0.1881	0.1381	0.1925	0.2146
Test scores	0.4055	0.0000	0.1644	NA	NA	8e-04	0.0103	0.0509
HS dropout	-0.3955	0.0000	0.1564	NA	NA	1e-04	0.036	NA
Seg poverty	-0.3892	0.0000	0.1515	NA	0.1868	0.0993	0.1926	NA
Seg income	-0.3749	0.0000	0.1406	NA	0.1775	0.0888	0.1835	NA
Manufacturing	-0.3535	0.0000	0.125	0.1354	0.1416	0.0776	0.1431	0.0268
Urban	-0.3526	0.0000	0.1244	NA	0.1244	0.1244	0.1244	NA
Seg affluence	-0.3526	0.0000	0.1243	NA	0.163	0.0761	0.1739	NA
Latitude	0.3511	0.0000	0.1233	0.1255	0.1256	0.1114	0.1259	0.1179
Seg racial	-0.3425	0.0000	0.1173	NA	0.1594	0.053	0.1658	NA
Local tax rate	0.3289	0.0000	0.1082	0.1148	0.1152	0.0778	0.1159	0.0978
Student teacher ratio	-0.3150	0.0000	0.0992	0.115	0.1076	0.0808	0.1102	0.1259
Longitude	-0.3061	0.0000	0.0937	NA	NA	0.0689	NA	0.1339
Violent crime	-0.2776	0.0000	0.0771	NA	0.1657	0.0041	0.1772	NA
Migration in	-0.2559	0.0000	0.0655	NA	0.1183	0.026	0.1498	NA
School spending	0.2392	0.0000	0.0572	0.0593	0.0586	0.0524	0.0589	0.0589
Chinese imports	-0.2086	0.0000	0.0435	NA	NA	0.0026	0.1571	NA
Progressivity	0.1883	0.0000	0.0355	NA	0.0428	0.0135	0.0422	NA
Colleges	0.1793	0.0000	0.0322	0.0644	0.0536	0.0053	0.0591	0.0261
Local gov spending	0.1749	0.0000	0.0306	0.0355	0.0349	0.0157	0.0355	0.031
Share01	-0.1625	0.0000	0.0264	0.0589	0.0443	0.0038	0.0498	0.0689
Migration out	-0.1534	0.0000	0.0235	NA	0.0363	0.0105	0.0434	NA
Labor force participation	0.1527	0.0000	0.0233	0.0214	0.0225	0.0245	0.0221	0.019
Population	-0.1347	0.0002	0.0181	0.2193	0.0923	6e-04	0.1374	0.0861
EITC	0.1195	0.0011	0.0143	NA	0.0224	0.0093	0.0253	NA
Graduation	0.0435	0.2369	0.0019	NA	NA	0.0022	0.0197	NA
Tuition	-0.0421	0.2524	0.0018	NA	6e-04	0.0024	3e-04	NA
Income	0.0335	0.3624	0.0011	0.0025	0.0017	3e-04	0.002	0.004
Foreign born	-0.0140	0.7036	2e-04	NA	0.003	2e-04	0.0052	NA

Table 2: Multivarite Summary Analysis

Predictor	Estimate	Std. Error	T-value	P-value	VIF
Tuition	0.0000000	0.0000004	-0.0831879	0.9337461	1.835598
Migration_in	-0.0881943	0.2763284	-0.3191649	0.7497777	5.423010
Local_gov_spending	0.0000010	0.0000028	0.3596044	0.7193438	2.245337
Urban	0.0015683	0.0035148	0.4461894	0.6557159	2.567701
Student_teacher_ratio	-0.0005020	0.0010213	-0.4915901	0.6232945	2.691779
Income	0.0000003	0.0000006	0.5111472	0.6095464	7.376884
Longitude	0.0001129	0.0002049	0.5512722	0.5817728	4.274026
Local_tax_rate	0.1329218	0.2378940	0.5587440	0.5766673	2.569307
Divorced	0.0796410	0.1417140	0.5619843	0.5744598	4.012991
School_spending	-0.0012863	0.0020664	-0.6224871	0.5339970	3.946428
Population	0.0000000	0.0000000	0.7186668	0.4727902	3.320686
Seg_affluence	-0.3178434	0.4163419	-0.7634191	0.4456897	161.727935
Social_capital	-0.0020212	0.0024304	-0.8316412	0.4061365	6.739760
Gini	2.9291051	2.8879163	1.0142624	0.3111063	41784.423510
Share01	-0.0293691	0.0288887	-1.0166300	0.3099797	11726.020668
Gini_99	-3.0327952	2.8880886	-1.0501047	0.2943409	22476.048718
Graduation	-0.0138569	0.0126421	-1.0960878	0.2737382	2.282166
Teenage_labor	-2.1254163	1.9275653	-1.1026430	0.2708838	6.087540
Chinese_imports	-0.0008122	0.0006989	-1.1620983	0.2459288	1.365216
Seg_income	1.0644737	0.8313054	1.2804844	0.2011600	525.197773
Married	-0.0891439	0.0670603	-1.3293093	0.1845477	6.407447
EITC	-0.0005897	0.0004092	-1.4410299	0.1504041	2.234908
Labor_force_participation	-0.0689459	0.0475633	-1.4495617	0.1480099	5.947273
Colleges	-0.1052767	0.0721909	-1.4583098	0.1455855	1.865790
Migration_out	-0.5249081	0.3379638	-1.5531488	0.1212244	5.010401
Test_scores	0.0004603	0.0002758	1.6691273	0.0959201	3.591586
Seg_poverty	-0.8582205	0.4470597	-1.9197000	0.0556482	130.147684
Middle_class	0.0864940	0.0426455	2.0282072	0.0432399	9.164688
Foreign_born	0.1070508	0.0498271	2.1484465	0.0323131	3.417968
Violent_crime	-3.1935957	1.4811365	-2.1561793	0.0317000	1.806381
(Intercept)	0.1766119	0.0722059	2.4459482	0.0149023	NA
HS_dropout	-0.1917857	0.0767866	-2.4976443	0.0129265	1.857290
Latitude	0.0014236	0.0005312	2.6800831	0.0076824	5.946518
Seg_racial	-0.0483740	0.0165440	-2.9239619	0.0036642	2.134496
Commute	0.0754846	0.0255139	2.9585711	0.0032852	8.003940
Black	0.0885645	0.0253982	3.4870325	0.0005458	10.527764
Single_mothers	-0.3469003	0.0833052	-4.1642116	0.0000387	16.377285
Progressivity	0.0056017	0.0011192	5.0050286	0.0000009	1.481558
Religious	0.0608218	0.0115715	5.2561513	0.0000002	2.294771
Manufacturing	-0.1727035	0.0252835	-6.8306915	0.0000000	2.745890


lm(Mobility ~ . − State − ID − Name)
Diagnostic Graphs Based on Table 2
Residuals vs Fitted
Residuals
0.15
−0.10
0.05
0.10
0.15
0.20
Fitted values
392
393
Diagnostic Graphs Based on Table 2
Q−Q Residuals
Standardized residuals
2
−4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
392
393
Diagnostic Graphs Based on Table 2
Scale−Location
√|Standardized residuals|
1.5
0.0
0.05
0.10
0.15
0.20
Fitted values
392
393
Diagnostic Graphs Based on Table 2
Residuals vs Leverage
Standardized residuals
2
−4
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
0.5
0.5
Cook's distance
393
629



Final Results

The strongest predictors of economic mobility, after applying transformations, are Single Motherhood ($R^2 = 0.5236$, Reciprocal), Black Population ($R^2 = 0.3587$, Cube-root), Commute ($R^2 = 0.3315$, Raw), Gini ($R^2 = 0.3078$, Reciprocal), and Middle Class ($R^2 = 0.285$, Squared). These variables consistently show strong relationships with mobility. Education-related factors (Test Scores, School Spending, Dropout Rate) remain weak predictors even after transformation, suggesting they play an indirect role rather than a primary one. Segregation measures (racial and income) negatively impact mobility, with transformations improving their explanatory power slightly. Government policy variables (Local Tax Rate, EITC, Spending) remain weak predictors, even with transformation, indicating that broader structural factors may have a greater influence on mobility than direct fiscal interventions. The results suggest that family structure, income inequality, and segregation play key roles in determining economic mobility, while education and policy interventions likely act as secondary influences.

Model Creation

Choosing Varibales

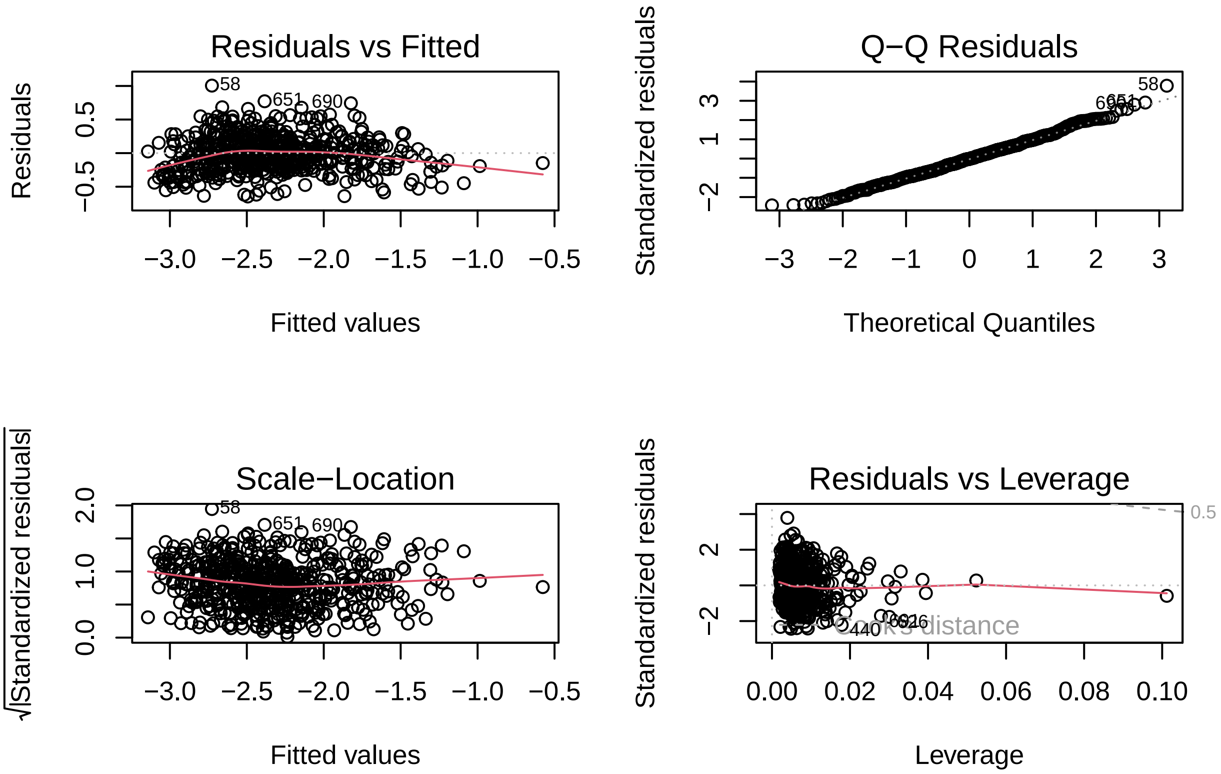
Criteria for our model require linear assumptions (Bi-Variate R^2 needs to be higher the 0.30), Co-Linearity has to be satisfied ($VIF < 5$), and heteroscedasticity has to me minimized to its max. Further we will be testing performance of model using a 20/80 test train split, and utilizing a StepAIC function to automate for our best model.

Table 3: Overall Data Model Summary Stats

Predictor	Estimate	Std_Error	t_value	P_value	VIF
(Intercept)	-4.1583603	0.0523081	-79.497450	0.0000000	NA
I(1/Single_mothers)	0.2209576	0.0129982	16.999150	0.0000000	2.038313
Commute	1.0043091	0.1049542	9.569026	0.0000000	1.476189
I(Middle_class^2)	0.7252797	0.1936497	3.745317	0.0001993	1.952914

Table 4: Overall Data Model Performance on Train and Test Data

Metric	Value
Train Adjusted R-squared	0.7004307
Test R-squared	0.7143003
Test RMSE	0.0305841
Test MAE	0.0209996



Model Creation Results

The Best Model that we created had three variables (Single Mothers, Commute, Middle Class). Further transformations made on these variables were the reciprocal of Single Mothers, and the square of Middle

Class. We also took the Log Odd of Mobility to normalize the Mobility values and improve our heteroscedasticity.

Extra choices we made were to impute to the mean on NA values since the NA values were low in these variables and we did not want to get rid of entire observations.

Addressing heteroscedasticity it has been minimized successfully and all plots across the board appear to be good looking. To Produce these plots all variables with a cooks distance greater than 4 were removed to achieve these results.

Testing Model on Regions

Here we then used the same model that was trained on all 80% of the train data. we then partitioned the test data out to be more granular down to each region. We then ran the model to see what its performance was on individual regions.

Table 5: Region-by-Region Performance (Test Set Only)

Region	Num_DP	Test_R2	Test_RMSE	Test_MAE
West	45	0.7939785	0.0318365	0.0233594
South	39	0.6926433	0.0264146	0.0186428
Northeast	17	0.2560313	0.0445602	0.0278136
Midwest	47	0.7466593	0.0260361	0.0182314

Regions Results

Based on the results it seems the model is doing alright in generalizing the data to all regions except for specifically the Northeast. The reason for this is probably due to low amounts of data points in Northeast causing there to be too much variance. Otherwise this model is pretty good at its prediction.

Key Takeaways & Recommendations

Based on our analysis, the data indicate that family structure (captured by the reciprocal of Single Mothers), commute times, and the strength of the middle class (using its squared transformation) are significant determinants of economic mobility. Our final model, which explains roughly 70% of the variability in mobility, shows that higher rates of single motherhood and greater income inequality are associated with reduced mobility, while shorter commute times and a robust middle class are linked to improved mobility. If more complete data were available for education-related factors, these variables might have contributed more strongly to the model; however, their high rate of missingness forced us to exclude them to avoid inflating variance. Similarly, while the model performs well overall, the lower predictive power observed in the Northeast—likely due to a smaller sample size—suggests that conclusions for that region should be interpreted with caution. In other words, if additional data were collected in underrepresented regions, the model's recommendations might be refined further. Overall, the evidence supports policy interventions that focus on enhancing family support systems, reducing income disparities, and improving transportation infrastructure to foster economic mobility. These conclusions are drawn as precisely as the current data allow, yet we acknowledge that further refinement of the model and data improvements could adjust these recommendations.