Socioeconomic and Demographic Predictors of Gas Station Density Across Counties in North Carolina (2020)

Sergiy Mouradkhanian, Max Dethlefs, Eleanor Curley, Juan Mateo Alvarez

Department of Economics, University of North Carolina at Chapel Hill

ECON 573.001 - Machine Learning and Econometrics

Professor Andrii Babii

April 13, 2025

Introduction

This study explores the extent to which socioeconomic, demographic, and infrastructural characteristics can predict the number of gas stations across counties in North Carolina. We tested our hypothesis that gas station density is dependent upon county-level factors such as income, education, population distribution, poverty levels, and the length of state highways. Gas stations are a critical component of transportation and economic access and their placement is unevenly distributed across the state. Identifying the fundamental patterns contributing to this can help identify where infrastructure gaps may be present, inform future planning decisions, and serve as a valuable tool for evaluating where new gas stations may be profitable.

This question holds practical importance for both policy and urban planning. In rural and underserved areas, inadequate access to fuel can constrain local economic development and limit mobility in the area. Although gas station distribution has been studied through spatial clustering and Geographic Information Systems (GIS) tools, few studies have applied predictive modeling. As a result, it remains unclear how effectively socioeconomic, demographic, and infrastructural variables alone can predict gas station presence.

Our contribution included applying machine learning models to evaluate whether a predictive relationship exists between county-level characteristics and gas station density. First produced was the baseline model with a report of its initial findings. A detailed description of the dataset follows. Variable selection techniques were applied as well as non-linear modeling approaches in order to complete the analysis of which model produced accurate and interpretable findings, improving the performance of the baseline model. Using data from 100 North Carolina counties and a set of 33 standardized and scaled predictors, the predictive models were applied as follows: linear regression, LASSO, Ridge, Elastic Net, principal components regression (PCA), generalized additive model (GAM), decision trees, and random forests. To ensure consistency as well as comparability, each model was trained using 10-fold cross-validation.

This analysis revealed that a cross-validation-selected linear subset model performed best overall. It achieved the lowest RMSE as well as selected a small set of four meaningful predictors which included poverty, infrastructure, and taxes. This highlights that gas station density can be predicted accurately with just a few key features. Linear regression resulted in a high R^2 but revealed issues with multicollinearity and heteroscedasticity. Ridge performed the worst and this is likely due to its inclusion of all variables. Elastic Net gave a balance of simplicity and inclusivity of variables. Nonlinear models, such as GAMs and random forests, captured some curved relationships, but did not significantly outperform simpler linear models. These results reveal that gas station density can be accurately predicted using a small set of socioeconomic, demographic and infrastructural factors, and that linear models are effective and interpretable in doing so.

Literature Review

Previous Research

Gas stations are a crucial component of urban infrastructure, as they directly influence transportation and the development of local economies. However, the density of gas stations are not equally distributed and this is due to a variety of socioeconomic, demographic, and infrastructural factors. Uncovering these factors is critical for urban planners and policymakers who aim to ensure equitable access to fuel as well as optimize urban planning and development. Some of the existing research on gas station distribution focuses on the relationship between socioeconomic factors, such as income, population density, and traffic patterns, and the density of gas stations in specific urban areas. For instance, J.J. Corn (1996) explored the history of how corporate strategies and marketing influenced gas station locations in the United States. Corn noted that wealthier and higher-density areas tended to have more gas stations present, driven by demand and corporate interests. This study highlights how economic factors as well as corporate strategies have a hand in determining gas station distribution. This offers important historical context but does not use predictive models to evaluate future gas station placement.

Similarly, Ende (2021) used t-tests and CART models to analyze the relationship between wealth-based indicators, such as income, home value, and net worth, and the density of branded gas stations. The study noted that higher-income areas were more likely to have branded stations, while lower-income neighborhoods were more likely to have unbranded stations. These findings suggest that socioeconomic factors have a strong influence on gas station distribution in terms of branding. More recent studies have utilized spatial clustering and GIS tools to map out where gas station are present in urban locations, such as Estelaji et al. (2023), who utilized GIS and

Weighted Linear Combination (WLC) models to locate the optimal spots for gas stations in Tehran. Their research incorporated both various socioeconomic and geospatial variables, and provided insights into how proximity to roads, population density, and economic factors influence the density and distribution of gas stations. While these studies provide useful insights, they are limited by their geographic focus and don't fully use predictive models to estimate gas station placement based on broader socioeconomic, demographic, and infrastructure patterns.

In contrast, Chen (2020) utilized county-level data and applied random forests and regression models to predict the location of EV charging stations, utilizing socioeconomic variables such as income, population density, and vehicle ownership. This predictive modeling approach was effective for this infrastructure placement and has not yet been applied to gas station density to our knowledge.

Contributions

This study addresses a significant gap in the literature by applying machine learning models, such as random forest and regression analysis, to predict gas station density across counties in North Carolina using 2020 data. While prior research focused on specific urban areas and spatial clustering, our study is unique in its urban-rural evaluation as well as use of machine learning and regression methods to evaluate which factors most strongly predict gas station densities. These models include linear regression, regularized linear models (LASSO, Ridge, and Elastic Net), PCA, GAM, decision trees, and random forests. Among these, the cross-validation-selected linear subset model emerged as most effective. It offered strong predictive accuracy along with strong interpretability due to its small set of variables. Random forests and other nonlinear models were also utilized to capture more non linear patterns. The machine learning models allowed for predictive analysis, which is useful for urban planners to assess gas station needs in relation to localized county-level conditions. By focusing on county-level data specific to North Carolina, this study provides a unique region-specific insight that differs from pure urban area focuses, like the Portland Metro area (Ende, 2021) or international examples (Zhu et al., 2024). These focuses, though important, do not fully capture rural-urban differences, such as those in North Carolina. In contrast to previous studies, using spatial clustering, descriptive methods, other infrastructure analysis, or urban specific focuses, this study introduces predictive models to offer actionable insights for urban planning and policy decisions within urban and rural settings. The model can help identify where gas stations are likely to be profitable or where there might be gaps in infrastructure. Therefore, these insights can be utilized as an important tool to improve infrastructure distribution and promote economic development.

Baseline Model Framework

We begin our analysis by applying an ordinary least squares (OLS) linear regression using all 33 available socioeconomic and demographic predictors to model the number of gas stations per county in North Carolina. This initial specification achieved an exceptionally high in-sample R^2 of 0.983 and an adjusted R^2 of 0.976, indicating that — at face value — these factors explain a large proportion of the variation in gas station density across counties.

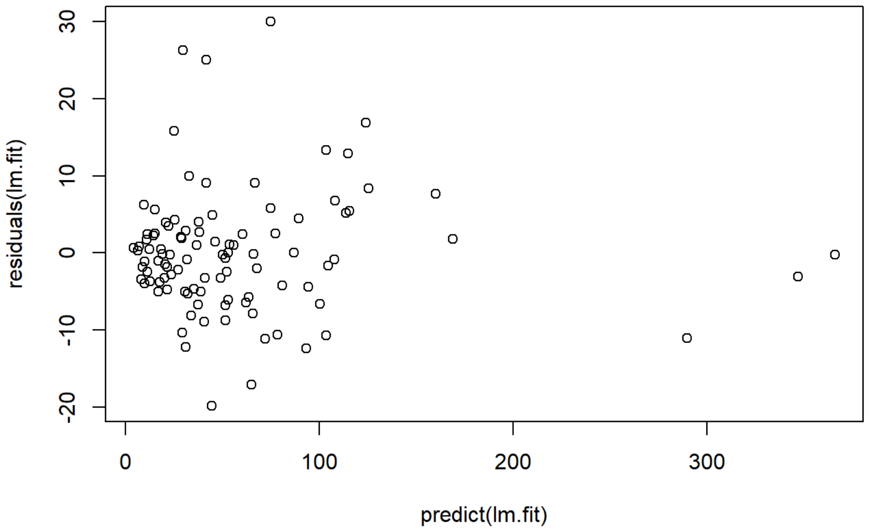

Figure 1. Residuals vs Fitted Values for Baseline Linear Regression Model

However, diagnostic checks immediately revealed key limitations of this baseline model. Our residual plot (Figure 1) displayed clear heteroscedasticity: the variance of residuals increased notably with higher fitted values, suggesting that the model struggled to predict gas station density accurately in larger, more urban counties. Additionally, despite the high R^2, many individual predictors were statistically insignificant, with large standard errors and low t-values, suggesting multicollinearity.

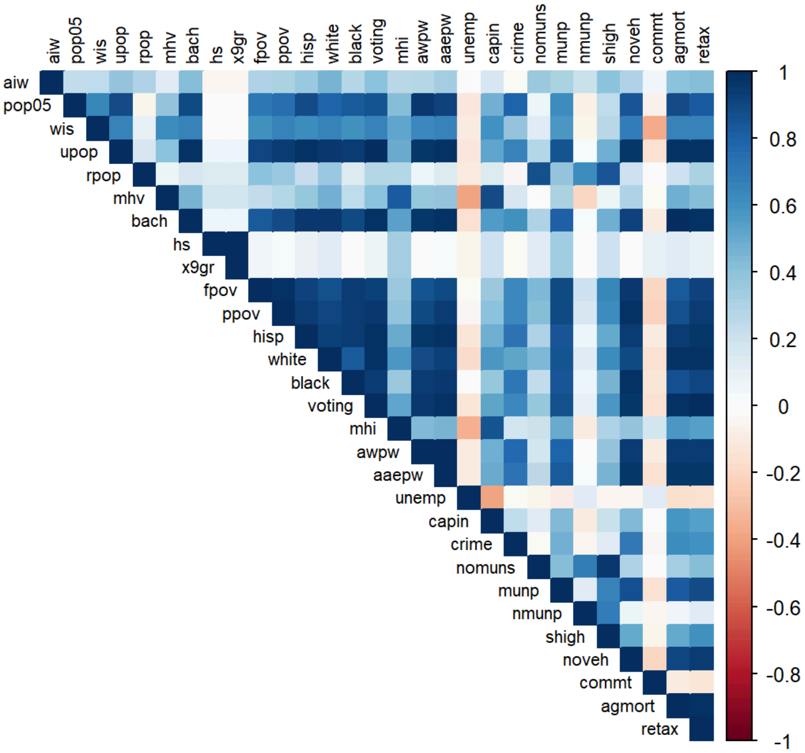

Figure 2. Correlation Matrix

Further investigation using a correlation matrix (Figure 2) and variance inflation factors (VIF) confirmed the presence of severe multicollinearity. Some VIF values exceeded 150,000, far beyond conventional thresholds of concern, indicating that many variables provided overlapping or redundant information.

In fact, four variables, total population (POP), percentage of other racial groups (OTHER), percentage of child population (CHILD), and median property value (PRPVAL), were automatically excluded from the model due to perfect multicollinearity. Their removal left the model's R^2 and adjusted R^2 virtually unchanged, further confirming that these variables did not add independent explanatory power.

To address the broader multicollinearity problem, we adopted a stepwise variable selection strategy guided by VIF values. This involved iteratively removing variables with the highest VIF scores, particularly when they were also statistically insignificant, and re-estimating the model at each step. This process gradually reduced the number of predictors from 33 to 8.

The final manually refined model retained variables that were both statistically meaningful and exhibited acceptable VIF values (generally below 5). These included: The percentage of Hispanic population (HISP), Total Crime Index (CRIME), Walkability Index Score (WIS), Percentage of rural population (RPOP), Miles of State Non-Municipal Secondary Paved Total (NMUNP), Air & Water Quality Index (AIW), Average Commute Time (COMMT), and the unemployment rate (UNEMP). Although this reduced model substantially mitigated multicollinearity, it came at the cost of lower explanatory power. The adjusted R^2 declined to 0.8956, and the residual standard error increased from 9.47 to 19.76.

To evaluate the predictive performance of these models beyond the training data, we implemented 10-fold cross-validation on each iteration. The results (Figure 2) indicate that while the fully saturated model achieved the highest in-sample fit, its cross-validated error was among the worst, a clear consequence of overfitting and instability driven by multicollinearity. Model performance improved as variables with extreme VIF values were removed. The model containing 24 predictors (iteration 4) achieved the lowest cross-validated error, suggesting that it struck the most effective balance between explanatory power and model simplicity. Further reductions in variables beyond this point increased prediction error slightly, suggesting potential underfitting.

Table 1

Initial Model and Application of Variance Inflation Factor

Linear Regression Model	**Predictors removed**	**Number of Predictors**	**Adjusted R²**	**Residual SE**	**RMSE**
Initial Model	-	33	0.976	9.474	2,088.88
After Removing Perfect Multicollinearity	POP, CHILD, OTHER, PPVAL	29	0.976	9.474	4,079.62
1st VIF-based Removal	HS, X9GR	27	0.9766	9.365	39.94
2nd VIF-based Removal	VOTING, UPOP, BACH	24	0.977	9.266	23.00
3rd VIF-based Removal	AWPW, RETAX, SHIGH, PPOV, AGMORT, NOVEH, WHITE, NOMUNS, FPOV, BLACK	14	0.9631	11.75	52.47

4th VIF-based Removal	AAEPW, CAPIN, MUNP, MHI, MHV	9	0.8961	19.72	42.75
5th VIF-based Removal	POP05	8	0.8956	19.76	47.13

While this manual, VIF-guided approach effectively addressed multicollinearity, it relied heavily on subjective decision-making and iterative variable removal. Recognizing these limitations, we next employ formal variable selection methods, including Best Subset Selection, Forward and Backward Stepwise Selection, and Principal Component Analysis (PCA), to complement and compare against our manual process before exploring regularization techniques such as Ridge and LASSO regression.

While our baseline linear regression analysis provided valuable initial insights into the relationships between socioeconomic factors and gas station density, it also revealed critical limitations, most notably, severe multicollinearity, heteroscedasticity, and instability in variable significance across model specifications. Before proceeding to formal variable selection techniques and model refinement, it is essential to provide a detailed overview of our dataset. The following section describes the variables included in our analysis, their definitions, sources, and key descriptive statistics, laying the groundwork for subsequent modeling decisions.

Data Description

The collected dataset contains comprehensive demographic and economic characteristics for all 100 counties of the state of North Carolina. Our data is cross-sectional for the year 2020. This specific year was chosen because it is the most recent complete data of gas station counts by county that was found. Thus, the predictors were aggregated for the same time frame to avoid time-specific effects. The unit of observation is the county, meaning each row in the dataset represents one county in North Carolina.

Table 2

Data Sources for Analysis

Data Sources		
Organization	**Variables Sourced**	**Timeframe of Data**
NC Department of Agriculture & Consumer Services	Gas Station Locations/Count	August 17th, 2020
US Census Bureau	County Population Estimates/Economic and Educational Characteristics/ Average Commute Times	2020
North Carolina Department of Transportation	County-specific Mileage (Highway and Road)	2021
North Carolina Association of County Commissioners	Home Prices	2020
Census Burea & American Community Survey	Poverty Statistics	2019-2023

For all predictors, data was found specifically on the county level. Gas station location data provided by the North Carolina Department of Agriculture and Consumer Services was also used to receive a by-county sum for our response variable. In the end, a complete set of one hundred observations was obtained, one for each of the counties in North Carolina, with thirty-five predictors in total. One of the limitations of the dataset is that there were two data sources not dated to 2020. This temporal mismatch could lead to overfitting for outdated patterns or bias our predictor selection. Furthermore, the plethora of predictors could present collinearity concerns, where predictors are highly correlated with each other or with time, leading to unstable or biased model coefficients. By implementing L1 regularization in the form of LASSO, the effects of collinearity were limited. Furthermore, the number of predictors was limited by using the best subset selection, forward selection, and backward selection in the linear model. One last concern regarding the data is the fact that only one hundred observations exist, which is partly due to the cross-sectional nature of the data and the limited number of counties in the state. However, since the number of predictors is less than half the number of observations, it remains in a viable realm.

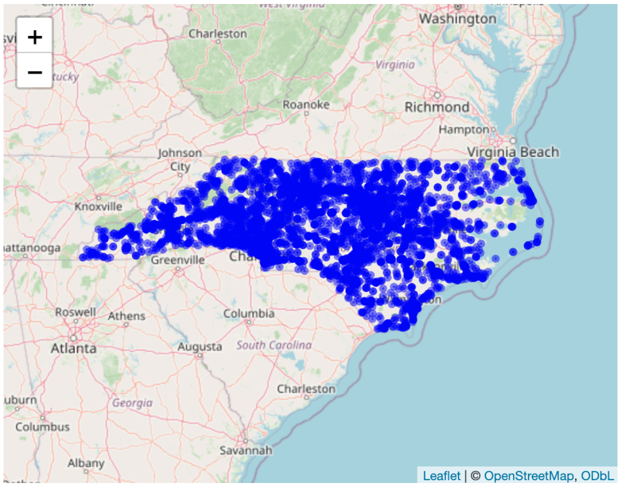

Figure 3. Map of gas station locations across North Carolina.

To supplement the county-level dataset, this map was created in R Studio using the leaflet package to visualize the latitude and longitude of each gas station recorded by the NC Department of Agriculture and Consumer Services. The plot highlights noticeable clustering around urban centers such as Raleigh and Charlotte. More rural regions, particularly in the far eastern and western parts of the state, show dispersed distribution. Though the main analysis uses county-level data, this visualization offers helpful context on the geographic spread of gas stations across the state.

Table 3

Listed Predictors with Summary Statistics

Abbreviation	Name	Mean	Median	Standard Deviation
GAS	Gas Station Count	5.80	4.25	6.12
WHITE	White Population	63,121	32,360	90,167
BLACK	Black Population	21,075	9,577	44,540
OTHER	Other Race Population	95,383	47,270	152,151
VOTING	Voting Age Population (18+)	81,551	39,646	132,107
CHILD	Childhood Population (<18)	22,843	10,855	40,011
MHI	Median Household Income	50,063	48,114	10,247
AWPW	Annual Wages by Place of Work	2,305,113,000	622,500,000	6,538,636,000
AAEPW	Average Annual Employment by Place of Work	41,864	15,244	94,065
UNEMP	Unemployment Rate (ACS)	4.57	4.32	1.32
CAPIN	Per Capita Money Income (Census)	27,696	26, 703	5, 278
CRIME	Total Index Crime	1,065	499	3,900
AIW	Air & Water Quality Index	1.94	0	4.57
POP05	Population Within 0.5 Miles of Public Transit	11,067	0.00	50,218
WIS	Walkability Index Score	5.94	6.00	1.46
UPOP	Urban Population	69,647	19,892	167,134
RPOP	Rural Population	34,747	29,354	21,734
MHV	Median House Value($)	158,571	149,400	57,092
BACH	Education: At least a bachelor's degree	25,214	8,136	60,222
HS	Education: Less than High School	256,939	4,568	2, 422,684

9GR	Education: Less than 9th Grade	112,987	1,530	1,073,057
FPOV	Families Below Poverty	2,525	1,645	3,122
PPOV	People Below Poverty	13,558	8,104	17,955
POP	Total Population	104,394	51,353	171,965
HISP	Hispanic Population	11,185	3,509	22,970
NOVEH	Households with no Motor Vehicle	2,231	1,089	3,664
SHIGH	Total State Highway Total (Miles)	766	721	415
NOMUNS	State Non-Municipal Primary Road System (Total)	588.94	527.33	312.80
MUNP	State Municipal System Total (Miles)	97.07	52.00	130.12
NMUNP	State Non-Municipal Secondary Paved Total (Miles)	116.39	110.27	58.00
COMMT	Average Commute Time (Minutes)	25.77	25.50	3.67
AGMORT	Aggregate Mortgage Value ($)	6,204,703,144	2, 747,533,950	12,247,294,378
PRPVAL	Median Property Value ($)	158,571	149,400	57,091
RETAX	Real Estate Taxes by Mortgage ($)	26,498	14,517	38,408

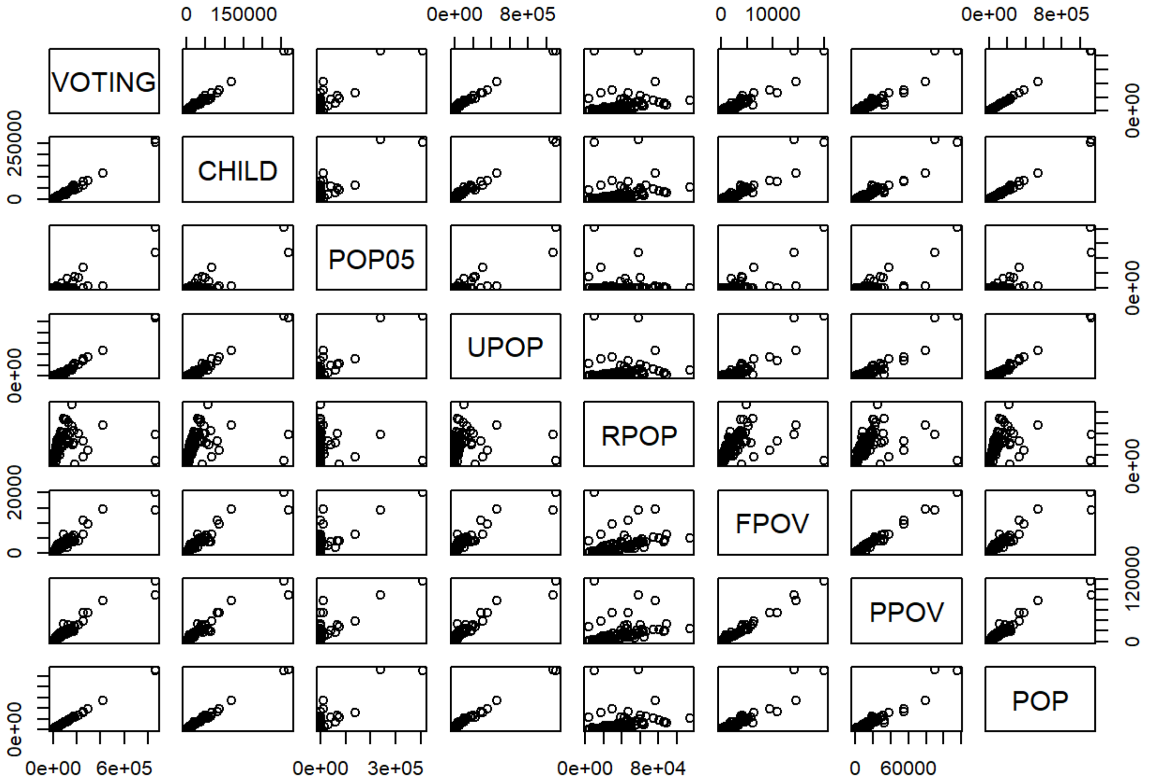

Figure 4. Variable Correlation Grid

Gas station density across counties is likely shaped by a combination of demand, economic conditions, and infrastructure. Counties with larger populations and more highway access are hypothesized to have more gas stations due to higher fuel demand and better accessibility. Economic factors like income, poverty rates, and property values may also play a role in influencing where businesses choose to locate. The core hypothesis is that socioeconomic and demographic variables, along with infrastructure indicators, can meaningfully predict the number of gas stations in a given county. Specifically, we expect that counties with higher populations, greater road mileage, and lower poverty rates would reflect a higher density of gas stations.

To model these relationships, the analysis uses a variation of supervised learning methods focused on prediction. The basic linear model is written as:

$$No.\ of\ Gas\ Stations_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$$

Where the response variable is the number of gas stations for a county i, and X_{ij}, represents the predictor value values for each predictor k = 1,..., 33, across all counties. This linear model is a starting point for more advanced methods, including LASSO, Ridge, Elastic Net, principal components regression, and random forests. These address nonlinear relationships, multicollinearity, and/or improve prediction accuracy based on the basic model.

With this basic understanding of our dataset, the task of model refinement follows. In the next section, several formal variable selection methods are applied, including Best Subset Selection, Forward and Backward Stepwise Selection, and Principal Component Analysis (PCA). This is implemented to systematically address the limitations of our baseline model and to identify a more parsimonious, stable, and interpretable set of predictors for gas station density.

Empirical/Econometric Model Framework

In order to test the validity of important predictors to the optimal model, a variety of econometric and statistical learning methods were implemented to analyze specific changes to selected variables in comparison to the baseline linear regression as evaluated in the Baseline Model Framework section. Multiple approaches were chosen to ensure the robustness of results (if the same factors emerge across methods, which would highlight the importance of specific predictors), and to highlight specific strengths across each approach (e.g., OLS for interpretability, LASSO for variable selection under multicollinearity, Random Forest for capturing non-linear interactions, etc.). The following methods were defined and implemented:

Model Selection

The best subset selection, forward selection, and backward selection were applied together because these methods aim to reduce the number of variables in a stepwise process based on criteria such as BIC for the optimal OLS model. Both backwards and forward selection were analyzed because each starts at different points in the original model (backward selection starts with the full model and eliminates the weakest variables while forward selection begins with zero predictors and adds the most important at each consecutive point). In doing so, it was hypothesized that each approach would capture the best possible variable given at different points regardless of direction, which would ultimately allow for the analysis of consistently chosen predictors despite slight variations in chosen variables. The best subset selection would further be used to validate forward and backward selection by defining the best predictors for the prediction model in relation to forward and backward selection.

LASSO, Ridge, and Elastic Net

The Least Absolute Shrinkage and Selection Operator (LASSO), or L1-regularization, is an important approach in variable selection as it shrinks coefficient values to zero, ultimately minimizing its significance to the model completely. LASSO was used in order to see a more "clear-cut" version of the variable selection process of gas station density predictors. It is an all-or-nothing approach that makes selection decisions between variables such as POP vs PPOV vs FPOV in order to simplify issues of collinearity completely. One weakness with extremely correlated data as referenced is the potential for random selection as a result of random minor data tweaks among the correlated variables. As a result, we chose to interpret the LASSO-selected predictors not necessarily as the most truthful model, but as a basis for defining the predictors that would be chosen most frequently throughout regularizations.

While LASSO shrinks irrelevant coefficients to zero, the Ridge approach (L2-regularization) penalizes these coefficients to values close to zero, but does not necessarily remove them. Instead of picking one as LASSO performs, the Ridge will still implement these variables and will instead distribute the effectiveness across. This is a useful approach because when it comes to various population measures that aim to define similar results, the Ridge approach will essentially equalize across multicollinearity. The penalty term (λ) as defined by cross-validation will result in choosing the variables that tend to have the greatest magnitude of coefficient values as the results.

As a basis of comparison to LASSO and Ridge in their own strengths, we decided to incorporate the elastic net as a measure that highlights both the strength of LASSO and Ridge simultaneously. In other words, the elastic net would allow us to shrink coefficients as well as choose the most important predictors, especially given multicollinearity. Overall, we predict the elastic net to be a more stable method when it comes to a large number of variables, as it would rely less on the data in comparison to an individual analysis by Ridge or LASSO.

Principal Component Analysis (PCA)

Instead of looking at the whole model with potentially correlated predictors, PCA aims to reduce the data to uncorrelated components. So instead of looking at individual variables, we aim to analyze how many dimensions or principal components explain the most variation across the predictors. We predict that the first principal component will be a “size of county” factor (due to its high loading on population-related variables) followed by “urban vs rural” or “high income vs low income” contrasts that aim to define splits within counties that contribute to high or low gas station counts.

In performing PCA, we would regress GAS on the first few principal components to see how much variance is attributed to each and whether or not there is a main factor that potentially accounts for most of the variability. We chose to perform PCA because it is useful in diagnosing near-collinearity through low eigenvalue which we would expect in our given predictors. This would also be accounted for in the methods mentioned above. We predicted that close to 90% of the variance in the predictors would be explained by the first three to six principal components given the extent to which collinearity exists. Essentially we would perform PCA regression analysis on a set of principal components to identify the highest R^2 value, and the more principal components that exist would imply other combinations of predictors such as rural-urban splits would account for gas station counts across counties.

Non-Linear Approaches

Generalized Additive Model (GAM)

The generalized additive model is very useful in identifying potential nonlinear relationships for particular predictors such as population and income which are continuous in nature. Using Splines, we would be able to define smooth nonlinear effects that would capture more effectively differences between counties on characteristics that we may not be able to find data for. For example, counties with larger cities may have larger gas stations that would fulfill the needs more efficiently compared with many smaller gas stations, so the relationship may not be proportional to factors such as population size and instead could factor diminishing returns.

Essentially, we focused on smoothing plots for the most important variables to generate effective relationships. For instance, if the GAM proposes relationships between population and gas count to be non-linear, then we may integrate alternate functions such as log population in order to improve the model's effectiveness. We predict that there could be possible diminishing returns to gas station counts given all other variables due to factors beyond statistical interpretation, hence why we decided to implement GAM into our analysis.

Decision Trees and Random Forests

We decided to implement a decision tree to see how the variables would get split based on the counties. Ultimately results from approaches like LASSO and linear regression would be used as a framework for which the decision tree would further validate when it comes to chosen variables. The tree would endogenously pick the most important variable to split first, indicating its importance. We predicted that the first split would come down to factors involving population size followed by splits in urban or rural factors. This would provide us with an idea of how the main differences between the counties can be attributed to each, ultimately grasping predictors to pinpoint.

While other approaches may capture linear relationships, the strength of decision trees come down to interactions between predictors as well as thresholds that create separations between counties. For instance, counties that have populations below fifty thousand may have less than a set number of gas stations, and those above would contain secondary or tertiary factors in play compared to other counties. In doing so, a tree would effectively determine piecewise constants that would help solve issues in nonlinearities.

In the application of the decision trees, random forests would allow us to achieve a more robust measure of the important variables as well as complex interactions that decision trees may not individually reveal. We may benefit from random forests as we obtain an importance score from each variable (measured as average MSE decrease) which will ultimately reveal the least informative variables in predicting gas station counts. Similar to decision trees, we anticipate population-based measures to be ranked higher in importance. However random forests may highlight certain variables within specific subsets of the data, capturing localized or conditional patterns that may not be evident when assessing variable importance across the entire dataset.

Justification of Models:

Each of the referenced models will be evaluated on their fit and prediction performance and be compared to the other methods for weaknesses. While the OLS with all variables may have a rather high R^2 value of 0.983, other more sparse models may have slightly higher adjusted R^2 values or lower BIC and Cp, indicating a more powerful prediction model. Additional metrics such as Root Mean Squared Error (RMSE) are important through cross-validation in order to define generalization capabilities. It is important to note that with only 100 observations, some estimates may be rather difficult to capture due to factors of instability. Hence, we have to place a stronger emphasis on sample-based R^2 values in reference with theoretical applications. The results section will present the outcomes of each method and define important values to help drive conclusions for the optimal model.

Results

Variable Selection

To complement our manual variable selection approach and provide a more systematic exploration of model refinement, we applied several formal variable selection methods: Best Subset Selection, Forward Stepwise Selection, Backward Stepwise Selection, and 10-fold Cross-Validation.

All three selection methods produced similar patterns when maximizing in-sample model fit. Specifically, the highest adjusted R^2 values were achieved with relatively large models, typically requiring between 12 to 16 variables depending on the method. Best Subset Selection, for instance, achieved its maximum adjusted R^2 of 0.9793 using 12 predictors, while Forward and Backward Stepwise Selection reached their peaks with 16 and 15 predictors, respectively.

However, once model complexity penalties were introduced, via Bayesian Information Criterion (BIC) and Mallows' Cp, the optimal model size shrunk considerably. Best Subset Selection, guided by BIC and Cp, favored an 8-variable model. Forward Stepwise's BIC minimized at 9 variables, while Backward Stepwise's BIC favored a 6-variable model.

Notably, several variables appeared repeatedly across these smaller, BIC- and Cp-selected models, particularly FPOV (female poverty rate), AWPW (average weekly pay), AAEPW (average annual earnings), and SHIGH (percentage with a high school degree). This consistency suggests their robustness as predictors of gas station density across North Carolina counties.

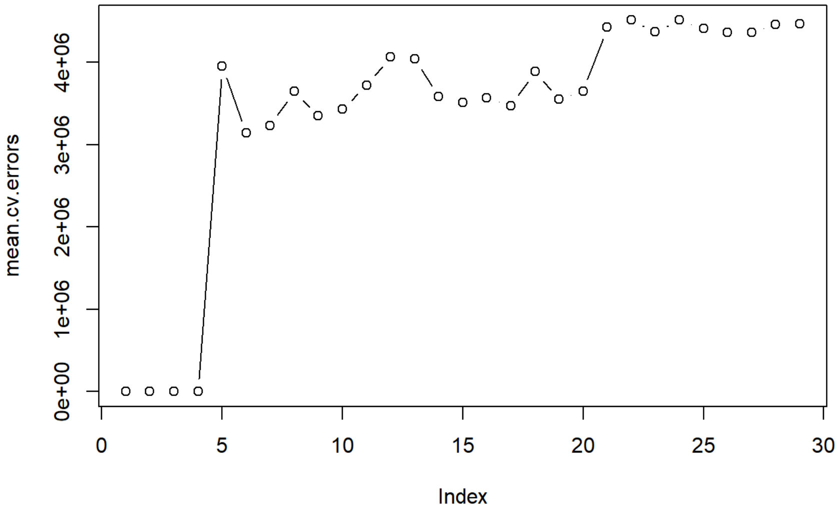

Figure 6. Mean Cross-Validation Error by Model Size from Best Subset Selection

To assess the true out-of-sample performance of these models, we applied 10-fold cross-validation to the Best Subset models of varying sizes. As shown in Figure 6, the cross-validation results revealed a clear pattern: test error decreased sharply up to a 4-variable model but increased substantially beyond that point. This indicates that adding variables beyond four led to overfitting and reduced predictive accuracy.

The optimal model identified by cross-validation contained just four variables: FPOV, AWPW, AAEPW, and SHIGH. Importantly, this aligns closely with the BIC-selected model from Best Subset Selection, providing further evidence of these variables' predictive strength.

Table 4

Summary of Optimal Models from Variable Selection Methods

Selection Method	**No. of predictors**	**Adjusted R^2**	**BIC**	**Cp**	**Variable Selected (summary)**
Best Subset	8 (BIC/Cp) / 12 (Adj R²)	0.9793 (12)	-353.13	-2.23	aiw, rpop, fpov, hisp, white, black, mhi, awpw, aaepw, shigh, agmort, retax
Best Subset	8	0.9789	-353.13	-2.23	aiw, upop, fpov, hisp, awpw, aaepw, shigh, retax
Forward Stepwise	9 (BIC) / 11 (Cp) / 16 (Adj R²)	0.9789 (16)	-342.60 (9)	2.71 (11)	rpop, fpov, ppov, hisp, white, awpw, aaepw, shigh, retax (+ others for 16 var model)
Backward Stepwise	6 (BIC) / 9 (Cp) / 15 (Adj R²)	0.9792 (15)	-349.86 (6)	0.47 (9)	rpop, fpov, hisp, awpw, aaepw, retax (+ others for 15 var model)
Cross-Validation	4	0.9743	-347.3519	11.61	fpov, awpw, aaepw, shigh

Note: Variables listed reflect those consistently selected across smaller, BIC- or Cp-favored models. Full variable lists for larger models included additional predictors.

Overall, while adjusted R^2 favored larger models, both information criteria (BIC and Cp) and cross-validation results strongly supported smaller, more parsimonious models. The consistent selection of FPOV, AWPW, AAEPW, and SHIGH across multiple methods and selection criteria suggests these four variables form a particularly stable and interpretable foundation for predicting gas station density across counties.

Regularization

To further refine our predictive models and address the challenges of multicollinearity and overfitting, we applied three regularization techniques: LASSO, Ridge Regression, and Elastic Net. Each method was implemented on a standardized version of the dataset using 10-fold cross-validation to optimize model parameters and evaluate predictive performance.

LASSO

To predict the number of gas stations in North Carolina counties using LASSO, 10-fold cross-validation was implemented on a scaled version of the dataset containing socioeconomic and demographic predictors. By using the LASSO approach, it was found that the best possible lambda value was 3.29. This yielded a root mean squared error (RMSE) of 11.16 and an R^2 of 0.966, which shows that the model explains a significant portion of the variance in gas station counts across counties in North Carolina.

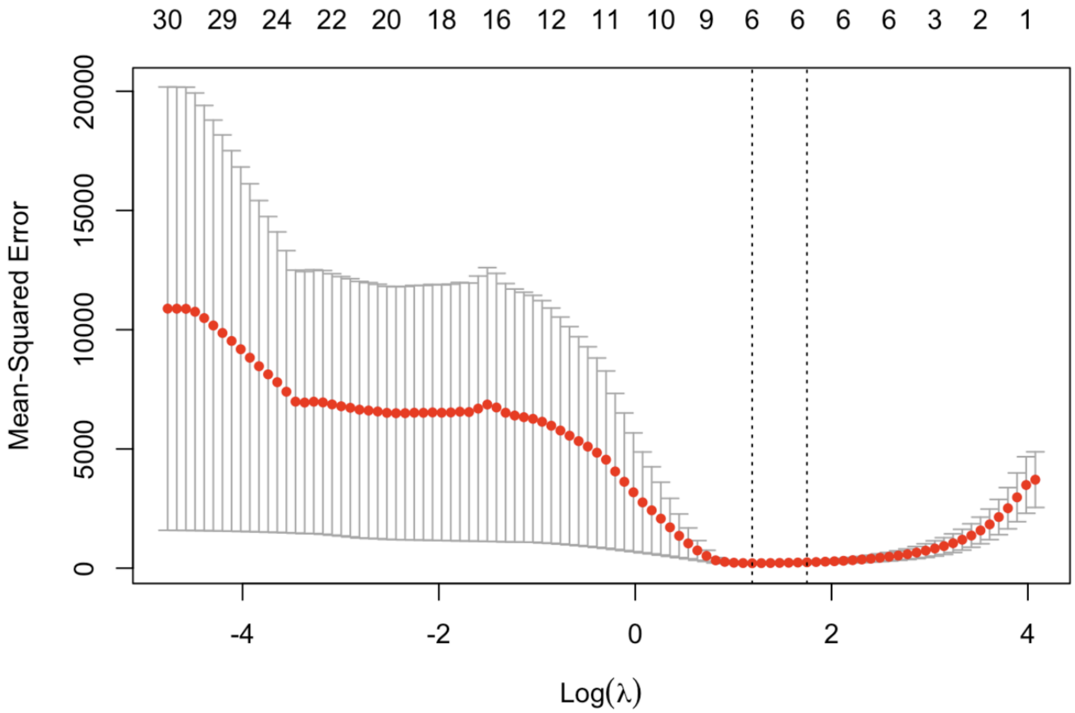

Figure 7. LASSO Mean-Squared Error For Respective λ Values in 10-Fold Cross-Validation

As shown in Figure 7, LASSO selected six variables as key predictors (2 dashed lines). Through R-code, the predictors identified were: families below poverty line (FPOV), people below poverty line (PPOV), white population percentage (WHITE), state municipal primary road length (MUMP), state highway length (SHIGH), and real estate taxes on mortgaged properties (RETAX). While the R^2 was rather high, the use of 10-fold cross-validation was set in place to mitigate risks of overfitting. Nonetheless, correlated variables (e.g., FPOV and PPOV) and the exclusion of some non-linear or interaction effects point to potential areas for improvement.

To emphasize potential improvements, three LASSO variations were tested. Adding an interaction between poverty and unemployment did not enhance model performance, so it was left out. Additionally, using polynomial terms generated a non-linear effect for variables such as state highway length, which slightly improved RMSE to 11.14 and raised R^2 to 0.9665. Finally, a log-transformed response variable (GAS) was used which resulted in different key predictors (walkability index score, families below poverty, and state highway length) and a lower R^2 of 0.722. This implies a reduced explanatory power compared with the original model. Overall, these results suggest the baseline or original LASSO model was the most balanced and tuned, with very small marginal gains from complexifying it.

Ridge

To compare with LASSO, a Ridge Regression model was fitted using the same standardized predictors and 10-fold cross-validation. Unlike LASSO, Ridge does not eliminate predictors completely. It is useful as it shrinks all coefficients toward zero to reduce model complexity and multicollinearity and to highlight the important predictors. By using the Ridge Regression approach it was determined the best possible lambda value was 743.07, resulting in an RMSE of 28.41 and an R^2 of 0.782. Compared to the LASSO model (RMSE = 11.16, R^2 = 0.966), Ridge performed much worse when it came to prediction accuracy and explained variance.

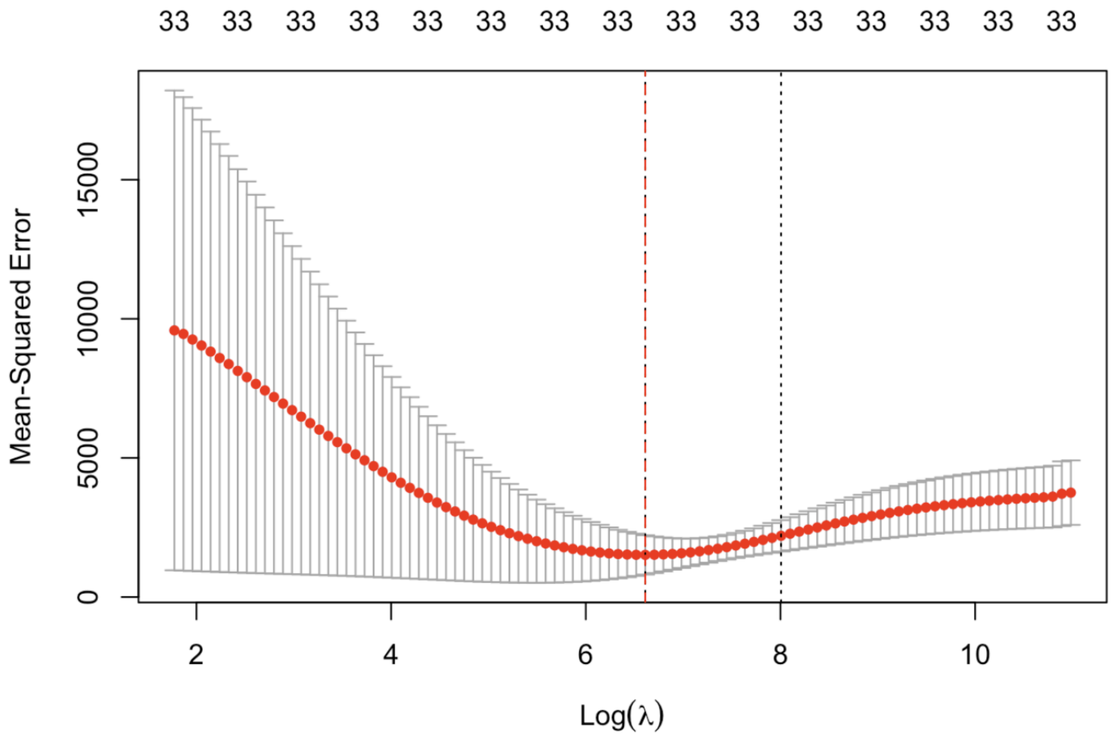

Figure 8. Ridge Regression Mean-Squared Error For Respective λ Values in 10-Fold Cross-Validation

As seen in Figure 8, the Ridge Regression retained all the predictors as seen by both dotted lines indicated at 33 predictors (total number of predictors). Through R-code, it was determined the largest coefficients were associated with: families below poverty (FPOV), people below poverty (PPOV), state municipal primary road length (MUMP), households with no motor vehicle (NOVEH), and real estate taxes (RETAX). Although Ridge included more information and insight by keeping all predictors, it did so at the cost of interpretability and performance. LASSO's sparsity not only produced a more concise model but also improved accuracy. We chose not to modify the Ridge regression model, as its baseline performance already lagged behind LASSO in both RMSE and R^2. Since Ridge maintained all predictors by design, adding interactions or nonlinear terms would likely complicate the interpretation of the model without meaningfully improving results. Instead of modifying Ridge, it was decided that alternative approaches such as the Elastic Net would offer a balance between Ridge and LASSO's respective strengths.

Elastic Net

To balance the sparsity of LASSO with the regularization strength of Ridge, an Elastic Net model was used with 10-fold cross-validation and an alpha of 0.5. This approach selected a broader set of predictors while still applying shrinkage to limit overfitting. By performing the elastic net, it was found that the optimal lambda was 11.51, which generated an RMSE of 13.04 and an R^2 of 0.954. These results fell between LASSO and Ridge in terms of accuracy and complexity.

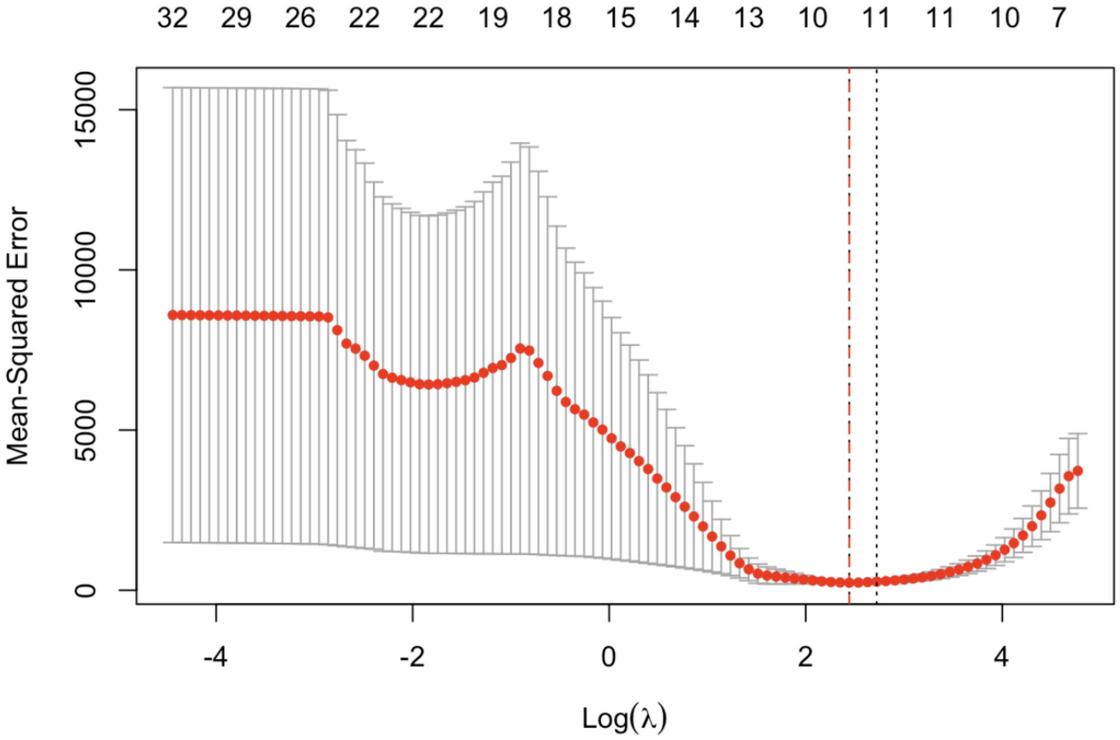

Figure 9. Elastic Net Mean-Squared Error For Respective λ Values in 10-Fold Cross-Validation

According to Figure 9, this approach maintained eleven predictors: families below poverty (FPOV), people below poverty (PPOV), total population (POP), white and other race population groups (WHITE) (OTHER), voting-age population (VOTING), childhood population (CHILD), municipal primary road length (MUMP), state highway length (SHIGH), households without motor vehicles (NOVEH), and real estate taxes (RETAX).

Compared to LASSO, Elastic Net included more demographic variables while still maintaining factors of exclusion for other less important factors. Although the slight increase in RMSE compared to that of LASSO suggests a lower accuracy prediction, the Elastic Net approach provided a subtly greater explanatory power through its greater emphasis on inclusivity. However, the model had to be evaluated on its predictive power, implying its shortcomings in comparison to the LASSO approach.

Table 5

Comparative Summary

Method	**Optimal λ**	**RMSE**	**R^2**	**Select Variables**
LASSO	3.29	11.16	0.966	fpov, ppov, white, mump, shigh, retax
Ridge	743.07	28.41	0.782	All variables retained
Elastic Net	11.51	13.04	0.954	11 variables including fpov, ppov, pop, white, shigh, mump, retax

Overall, LASSO produced the most accurate and interpretable model, balancing sparsity with predictive performance. Ridge Regression retained all predictors but performed poorly in comparison. Elastic Net offered a compromise between these two extremes, including more variables than LASSO but achieving slightly lower accuracy. Given our focus on predictive performance and model interpretability, the LASSO model emerges as the most effective regularization approach for modeling gas station density in North Carolina counties.

PCA

Another important approach in highlighting the most important variables revolved around analyzing the principal components. Ultimately any correlation between predictors were factored out, leaving each predictor as its independent entity in comparison to others. By performing PCA using R-code, it was found that the first four principal components accounted for 85.561% of the total variance. The following proportions of variance were found for the first five principal components in order respectively: 0.596979, 0.1088, 0.09498 0.05494, and 0.03552. This effectively implies that the data has a lower-dimensional structure despite having many variables. It is important to note that with each consecutive principle component, the proportion of variance decreased indicating weaker impact with more variables included.

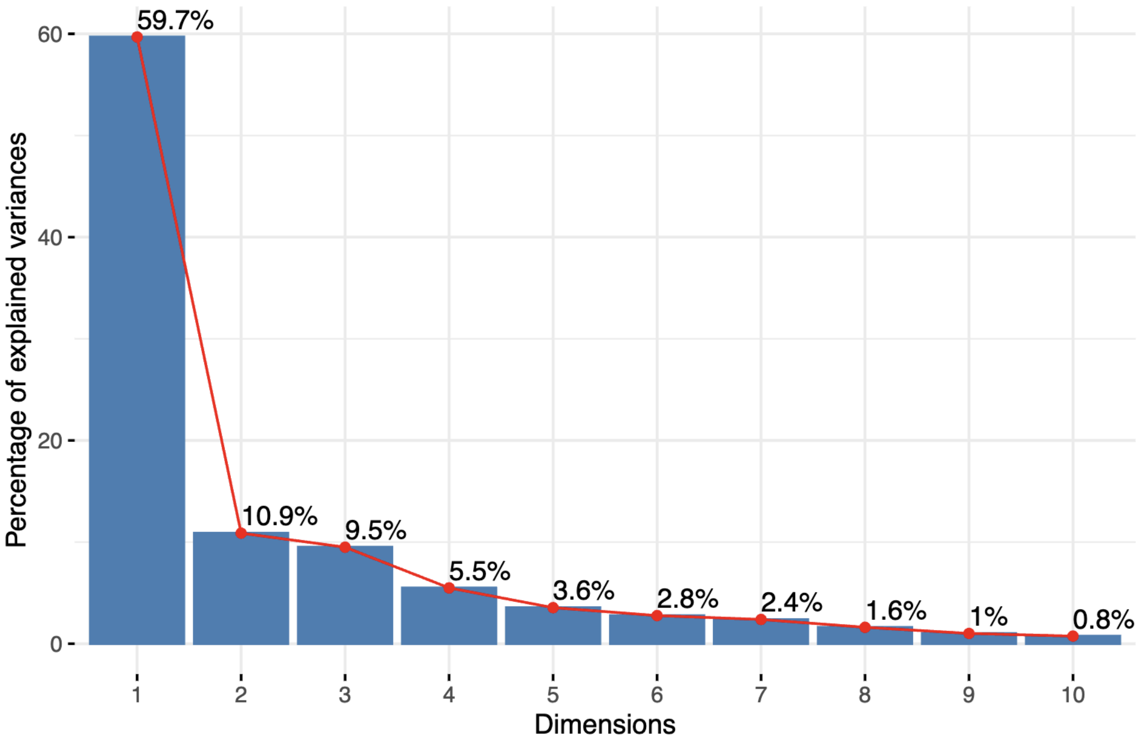

Figure 10. Screen Plot of PCA Highlighting Variance With Additions of Consecutive Principle Components

Based on Figure 10, it can be implied that the greatest variance tradeoff was predicted between the first 5 principle components after which the difference in variances became consecutively more stable. In order to test this prediction, a model was constructed for three, four, five, six, and seven principle components respectively such that the R^2 value was analyzed and compared. Hence, by performing a summary of the models, it was found that the R^2 increased from three to five principal components, but then began to decrease at six and seven components. As a result, the model with the highest R^2 was 0.9509 with five principal components. However it is important to understand that although the explanatory power may increase slightly with the increase of principal components, they contributed to less than 1% of the difference. This indicates there are minor roles from less effective variables that may contribute to the predictions potentially adding to the complexity of the model. As a result, we decided that the optimal measure contained two principal components as any further complexity would begin to deteriorate the interpretability of the analysis.

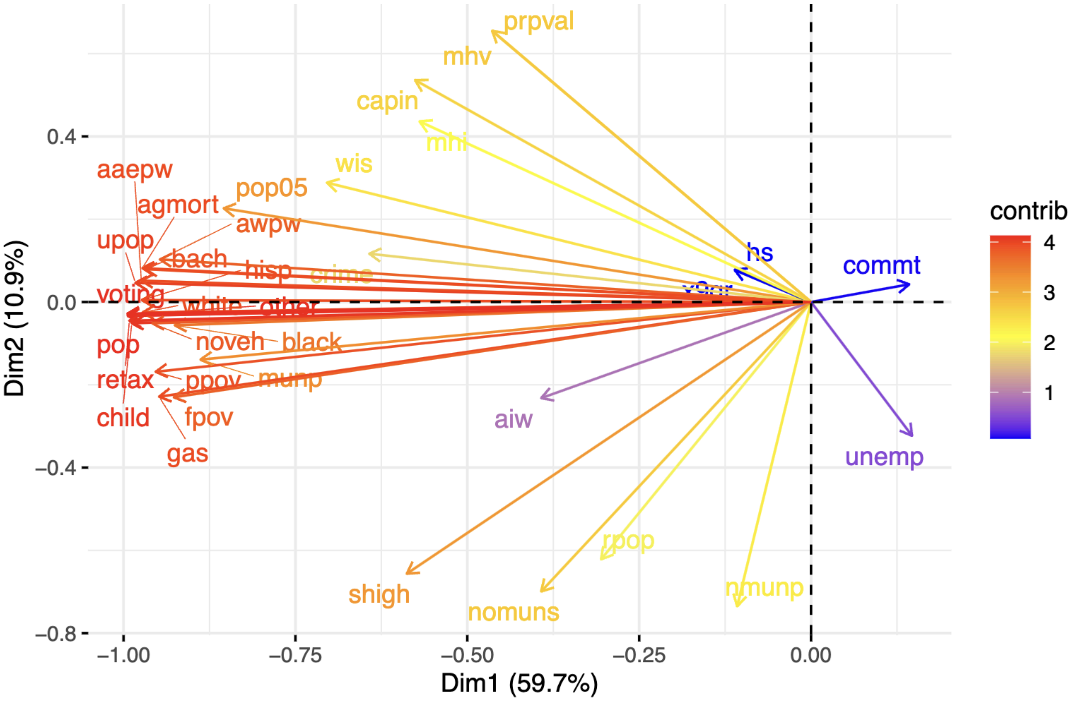

Figure 11. Visual of Variable Loading on PC1 and PC2

The loadings on PC1 were all rather uniformly spread across all predictors. In fact, every single predictor, except for a few, had loadings of approximately −0.15 to −0.25 on PC1. Although negative in sign, this uniform loading indicates that PC1 captured a general "size of county” factor. Specifically, counties that scored very negatively (such as Mecklenburg County) were large in terms of population, economic activity, and overall scale, whereas counties with higher scores on PC1 were smaller and less economically active. The negative sign of these loadings are not really as effective because these measures are arbitrary and can be reversed without affecting meaning.

The second principal component, while only contributing 11% of the variance, was able to capture a contrast between urban and rural counties/communities. The highest loadings on PC2 were NMUNP (0.346), NOMUNS, SHIGH (0.346), RPOP (0.325), and UNEMP (0.170). Such variables were indicative of more rural and less developed characteristics. On the other hand, the loadings with the most negative values included MHV / PRPVAL (-0.342), CAPIN (-0.279), and MHI (-0.227). Interestingly, these variables are indicative of characteristics pertaining to regions with more urbanization and wealth.

Essentially, PC1 and PC2 capture factors that encompass how large and populous a county is and its classification under urban or rural provided the size of the county. We were able to conclude that the PCA approach was important in reinforcing insight of regression approaches. It highlighted that the variations in the number of gas stations per county were largely due to differences across population sizes and their respective placement as rural lower income areas or urban wealthier areas.

Exploring Non-linear Modeling Approaches

To further investigate whether non-linear relationships could improve predictive performance, we applied several flexible modeling techniques to the variables identified in our best-performing linear models. Specifically, we focused on comparing the LASSO-selected model (6 variables) and the Cross-Validated Best Subset model (4 variables) under non-linear frameworks, including Natural Splines, Generalized Additive Models (GAM), Regression Trees, and Random Forests.

Initial Linear Comparison

OLS regressions using both the LASSO and Best Subset variables provided strong baselines for comparison. However, we observed several high-leverage urban counties with extremely high gas station counts, suggesting potential non-linearities in the data.

Out-of-sample cross-validation results indicated that the smaller Best Subset model achieved the lowest prediction error (RMSE = 10.49), outperforming both the LASSO model (RMSE = 13.85) and the earlier VIF-based model (RMSE = 23). While the LASSO model included more variables, the simpler Best Subset model demonstrated superior predictive accuracy.

Table 6

Comparative Performance Summary

Model	**Model Size**	**Adjusted R²**	**Residual SE**	**AIC**	**RMSE**
VIF-Based Selection	24	0.977	9.266	-	23.00
LASSO	6	0.9681	10.92	770.7408	13.85
CV Best Subset	4	0.9743	9.798	747.092	10.49

Given the potential presence of non-linear relationships in several variables, we first explore their bivariate relationships with gas station density using simple scatterplots

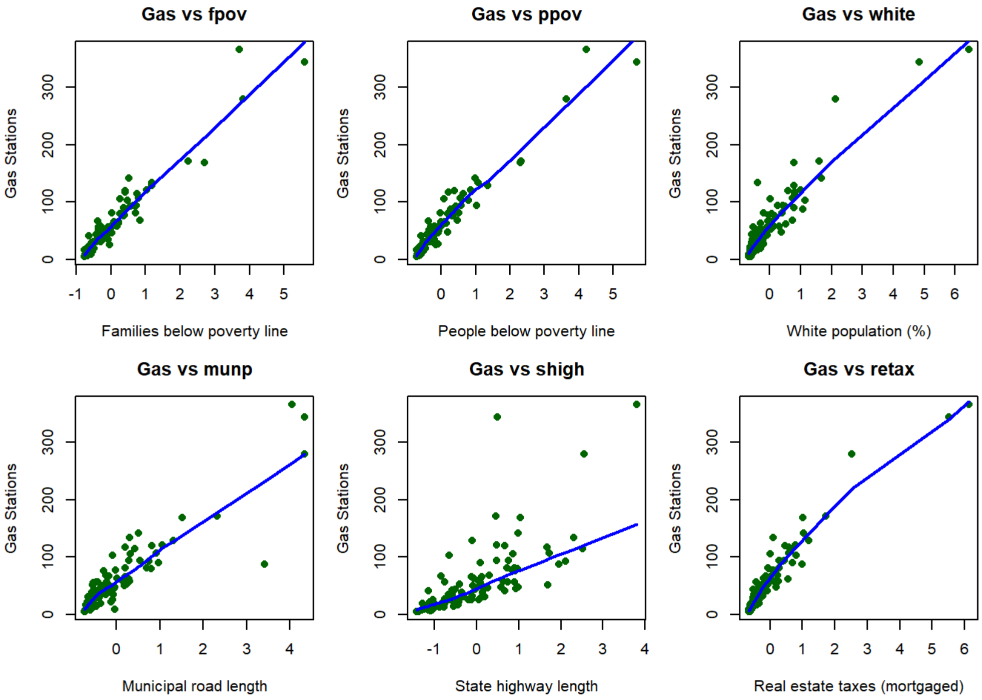

Figure 12. Scatterplots of gas stations versus selected predictors from the LASSO and Best Subset models.

Visual inspection suggests that variables such as AWPW, RETAX, and white population percentage exhibit patterns inconsistent with a purely linear relationship — particularly at higher values. To more formally capture these effects, we next apply flexible non-linear models, including Natural Splines and GAMs.

Natural Splines

To capture potential non-linear effects in these predictors, we fit Natural Spline models with 4 degrees of freedom to both the LASSO and Best Subset variable sets. Results indicated modest improvements in in-sample fit over linear models. Notably, variables such as real estate taxes (retax), white population percentage (white), and average wages (awpw) exhibited significant non-linear effects.

Table 7

Natural Splines Comparative Summary

Model	Adjusted R^2	Residual SE	AIC	CV RMSE
LASSO + Splines	0.9714	10.33	774.12	14.69
Best Subset + Splines	0.9753	9.61	753.61	13.72

While splines provided superior in-sample performance, cross-validation indicated higher test error than the purely linear Best Subset model, highlighting the trade-off between flexibility and overfitting in smaller samples.

Generalized Additive Models (GAM)

We next employed GAMs, which allow for variable-specific non-linearity while maintaining interpretability. For the LASSO model, GAMs achieved an adjusted R^2 of 0.978, while the Best Subset GAM achieved 0.976. The ability to flexibly model variables like white population percentage and municipal road length proved valuable.

Interestingly, in the LASSO GAM, the variable ppov was found to be statistically insignificant once fpov was included — likely due to their high correlation. Removing ppov improved model parsimony without sacrificing performance.

Table 8

Comparative Performance Summary

Model	**Adjusted R^2**	**AIC**	**CV RMSE**
LASSO GAM	0.978	742.29	25.19
LASSO GAM (without ppov)	0.978	741.32	19.75
Best Subset GAM	0.976	742.08	10.76

GAMs offered a significant improvement in modeling flexibility, though the Best Subset GAMs remained the most accurate and parsimonious model overall.

Regression Trees

Regression trees provided a highly interpretable segmentation of the data but were less accurate compared to other methods. In both the LASSO and Best Subset models, fpov emerged as the sole splitting variable — emphasizing its dominant role in predicting gas station density. While simple and easy to interpret, the tree's exclusion of other predictors limited its accuracy.

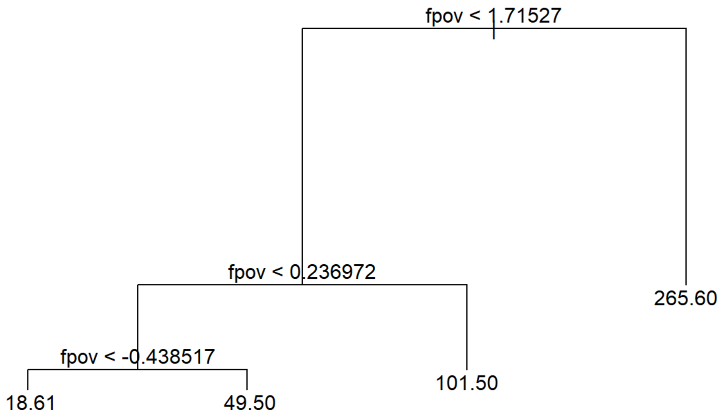

Figure 13. Regression tree model predicting gas station density.

The tree identifies the percentage of families below the poverty line (fpov) as the sole splitting variable, creating four distinct poverty thresholds. Counties with higher fpov values are consistently associated with greater numbers of gas stations.

Random Forests

To capture complex interactions and non-linear patterns missed by single trees, we applied Random Forests. The ensemble approach achieved competitive performance (RMSE = 8.85) and explained 89.9% of the variance in gas station counts.

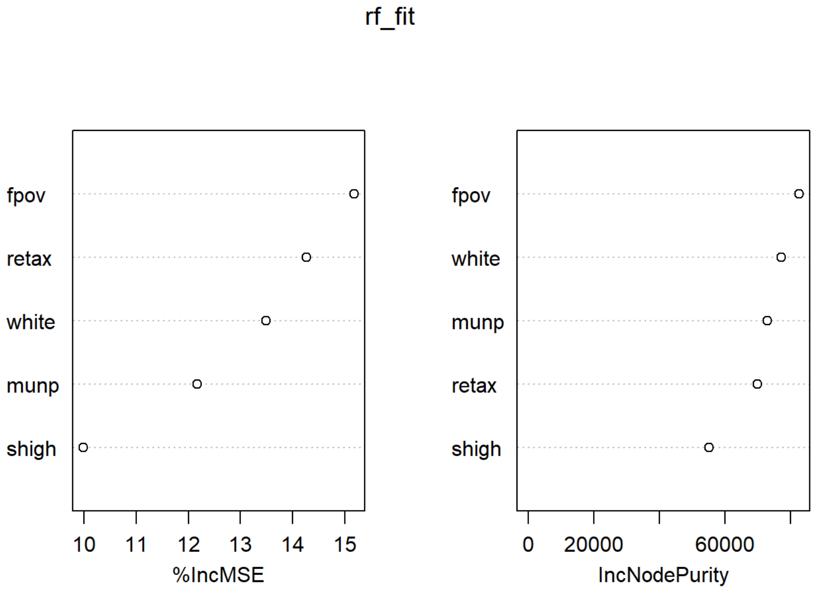

Figure 14. Figure X: Variable importance plot from Random Forest model, based on percentage increase in mean squared error (%IncMSE) and increase in node purity (IncNodePurity).

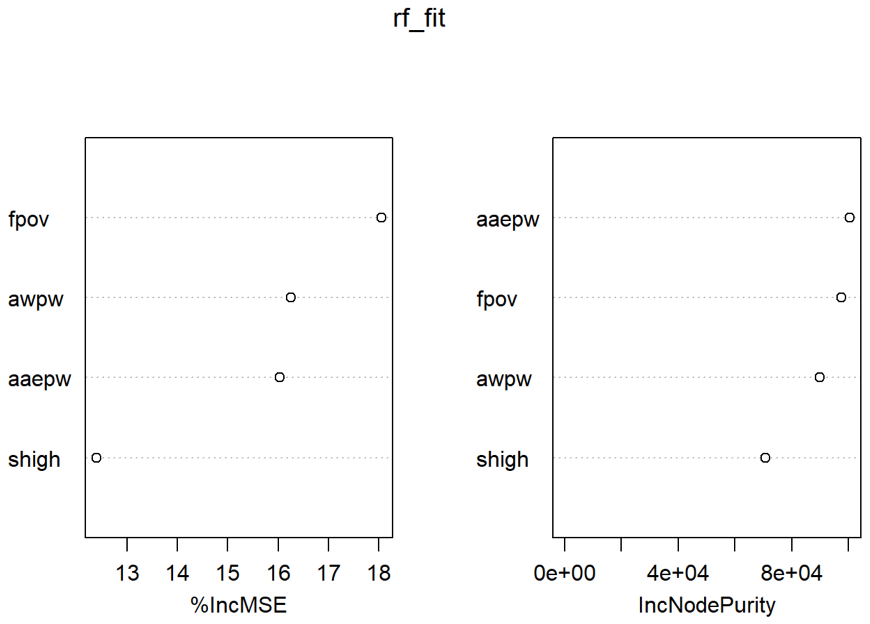

Figure 15. Variable importance plot from the Random Forest model using the Cross-Validated Best Subset variables (FPOV, AWPW, AAEPW, SHIGH).

Variable importance is measured by the percentage increase in mean squared error (%IncMSE) when each variable is removed (left) and the increase in node purity (IncNodePurity) from splits using each variable (right). Variable importance measures confirmed FPOV as the dominant predictor but also highlighted the significance of retax, white population percentage, municipal road length, and average annual employment (AAEP2) — consistent with previous models

Table 9

Comparative Performance Summary

Model	**Adjusted R²**	**CV RMSE**
LASSO GAM (without ppov)	0.978	19.75
Best Subset GAM	0.976	10.76
Random Forest	0.899 (variance explained)	8.85
Regression Tree	-	22.56

Overall, the GAM applied to the Best Subset model variables achieved the best combination of accuracy, flexibility, and interpretability. Random Forests slightly outperform GAMs in terms of raw predictive accuracy but at the cost of model transparency. Regression trees were highly interpretable but less accurate, while Natural Splines improved in-sample fit but struggled to outperform linear models in out-of-sample prediction. These results reinforce the robustness of the Best Subset-selected variables (FPOV, AWPW, AAEPW, SHIGH) while highlighting the presence of important non-linear effects in several key predictors.

Conclusion

This study assessed how socioeconomic, demographic, and infrastructural variables can predict gas station densities across 100 North Carolina counties using different machine learning models. 10-fold cross-validation and 33 variables were utilized to compare the performance of models based on both interpretability and out-of-sample predictive accuracy. The linear regression model performed well in terms of in-sample fit, with an adjusted R^2 of 0.976 and an RMSE of 9.47. However, residual analysis highlighted issues with multicollinearity and heteroscedasticity. This was particularly evident in urban counties that contained higher gas station counts. Therefore our baseline model was limited due to its weaker reliability for out-of-sample prediction. The best subset model selected via cross-validation produced the lowest out-of-sample RMSE of 10.49 and an R^2 of 0.95, using four variables. This highlights that a small, meaningful set of predictors can perform better than more complex models when it comes to predictive accuracy and interpretability.

The LASSO regression model selected six predictors (poverty levels, white population percentage, municipal road length, state highway length, real estate taxes, and people below poverty line) and achieved an out-of-sample RMSE of 13.85 and an R^2 of 0.966. It balanced accuracy and interpretability best among regularized methods. It was, however, slightly less accurate than the cross-validation-selected subset. Ridge regression kept all predictors and performed the worst among regularized models, with an RMSE of 28.41 and an R^2 of 0.782. This is likely due to its inability to eliminate noise effectively in retaining all variables. Elastic Net was a compromise, selecting 11 predictors and producing an RMSE of 13.04 and an R^2 of 0.954. While Elastic Net offered slightly more accurate predictions than LASSO, LASSO explained more variance with its R^2 and remained the most concise model among the three.

Principal components regression revealed the presence of two dominant components, one representing overall population size and another representing an urban-rural income divide. The interpretability, however, was more limited due to the abstract nature of principal components. Nonlinear models offered additional improvements in terms of flexibility and performance. Generalized additive models (GAMs), applied to the four predictors in the best subset model, highlighted nonlinear patterns amongst variables, such as white population percentage, municipal road length, and average wages. Although CV RMSE values appeared high, GAMs achieved some of the highest adjusted R^2 values, being up to 0.978. The GAMs applied to the best subset variables offered the strongest balance of accuracy, flexibility, and interpretability. Random forests slightly outperform GAMs in terms of predictive accuracy, with an RMSE of 8.85, but was less interpretable. Regression trees were highly interpretable but less accurate with an RMSE of 22.56. These findings highlight the effectiveness of the best subset-selected variables (FPOV, AWPW, AAEPW, SHIGH), while also highlighting the potential presence of important nonlinear effects in several predictors. Future research may benefit from stratifying counties by urban-rural classification or incorporating more granular, spatially explicit data to further improve forecasting across diverse regions.

In conclusion, this project evaluated whether socioeconomic, demographic, and infrastructural variables could be used to predict gas station density across North Carolina counties. While several models performed well, the cross-validation-selected linear subset model stood out for its balance of accuracy, with an RMSE of 10.49, simplicity, and interpretability. Its success highlights that a small set of meaningful variables can capture much of the variation in gas station distribution. This insight holds value for infrastructure planning across urban and rural areas. By shifting from descriptive mapping to predictive modeling, our approach offers a unique tool for identifying where new gas stations may be most needed or profitable. This work adds to the existing literature by showing that predictive models, when paired with public data, can help guide more equitable and data-driven decisions in infrastructure and economic development across rural and urban areas.

References

Corn, J. J. (1996). Review of *The gas station in America*, by J. A. Jakle & K. A. Sculle. *Winterthur Portfolio, 31*(1), 92–98. http://www.jstor.org/stable/4618539

Ende, J. (2021). *Gas stations and the wealth divide: Analyzing spatial correlations between wealth and fuel branding* (Publication No. 2528848513) [Doctoral dissertation, ProQuest Dissertations & Theses Global]. https://www.proquest.com/dissertations-theses/gas-stations-wealth-divide-analyzing-spatial/docview/2528848513/se-2

Estelaji, F., Naseri, A., Keshavarzzadeh, M., Zahedi, R., Yousefi, H., & Ahmadi, A. (2023). Potential measurement and spatial priorities determination for gas station construction using WLC and GIS. *Future Technology, 2*(4), 24–32. https://fupubco.com/futech/article/view/98

Roy, A., & Law, M. (2022, June 1). Examining spatial disparities in electric vehicle charging station placements using machine learning. *Sustainable Cities and Society.* https://www.sciencedirect.com/science/article/pii/S2210670722002980