Code

Predicting Median House Value in Philadelphia

Akira Di Sandro, Sofia Fasullo, Amy Solano

2024-10-14

Introduction

This study looks to understand whether several demographic and housing factors can be used to reasonably predict the median house value of a given census block group area in the city of Philadelphia, Pennsylvania. These variables include the percent of houses that are vacant on the block (`PCTVACANT`), the percent of residents with at least a bachelor's degree (`PCTBACHMOR`), the number of households living in poverty (`NBELPOV100`), and the percent of homes that are single-family detached units (`PCTSINGLES`). We used 2000 Decennial Census data for this study (Bureau 2000).

With each of the following variables that we will regress median household value upon, a correlation between them and the median household value can be inferred:

- The presence of vacant buildings on a block would decrease the value of the houses around it, and the vacant buildings themselves would be a lower value. Therefore, the percentage of vacant houses would be negatively correlated with median house value.
- Typically, it is assumed that people with bachelor's degrees are more likely to land higher-paying jobs than people without, and therefore can afford more expensive homes. The percentage of people with bachelor degrees would be positively correlated with median house value.
- If a household is living below the poverty line, they are unlikely to be able to afford expensive living. Therefore the number of households living below the poverty line would be negatively correlated with median house value.
- Owning a single-family detached house is culturally considered "the American dream," because one does not have to share walls or amenities with other people. The percentage of single-family units would be positively correlated with the median house value.

Methods

Data Cleaning

The original census data set had 1816 block groups included, also referred to as observations or rows. It was cleaned prior to this analysis so that it only contained 1720 observations. In the cleaning process of this data set, block groups were removed for one of the following four reasons: (1) total population < 40; (2) no housing units; (3) median house value is lower than $10,000; and (4) One North Philadelphia block group which had a very high median house value (over $800,000) and a very low median household income (less than $8,000). We move forward with our analyses using the cleaned version with 1720 observations.

Exploratory Data Analysis

Before regressing median house value upon any of the predictors, we want to make sure the assumptions held for a model are correct. These include the normality, homoscedasticity and non-collinearity of variables. This will be examined through exploratory analysis. We will examine the summary statistics and distributions of the variables to check some of these assumptions.

In addition to assessing the distributions of the variables, we will check for correlation between the predictor variables. When regressing median household value on other variables using an OLS linear model, we are assuming that the predictor variables are not highly correlated with each other.

Correlation is the assessment of whether or not a linear relationship exists between two variables. Correlation is standardized and normalized covariance, a statistic that measures the same thing. As such, the correlation between two variables falls between -1 and 1, with the values on either end being the strongest negative and positive linear relationship, respectively, and the value 0 being no evident linear relationship. Typically, statisticians consider an absolute value of 0.8 or higher to be a high correlation value, however this varies depending on the field of application.

The formula for correlation is as such:

$$r = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

where r is the Pearson Correlation, n is the number of observations in the sample, x is one variable, y is the other variable, x_i and y_i are the i^{th} observation of x and y, respectively, and $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively, and s_x and s_y are the sample standard deviations of x and y, respectively.

Multiple Regression Analysis

We use ordinary least squares (OLS) regression for our study. An OLS regression allows us to explore the relationship between our dependent variable (in this case, Median House Value) and possible explanatory variables (also known as predictors). OLS regression gives us information about the strength, direction, and statistical significance of the relationships between our each of predictors and the dependent variable. An OLS regression also gives us the ability to calculate how much the dependent variable will change when we change the value of one of the predictors while keeping all other variables constant. In our study, we will be using a multiple regression – an OLS regression with multiple predictors.

$$\text{MEDHVAL} = \beta_0 + \beta_1 \text{PCTBACHMOR} + \beta_2 \text{PCTVACANT} + \beta_3 \text{PCTSINGLE} + \beta_4 \text{NBELPOV100} + \epsilon$$

where β_0 is the intercept, or the value of the dependent variable when all predictors are set to zero, β_1 is the correlation coefficient for `PCTBACHMOR`, β_2 is the correlation coefficient for `PCTVACANT`, β_3 is the correlation coefficient for `PCTSINGLE`, β_4 is the correlation coefficient for `NBELPOV100`, and ϵ is the error

or residual term, which is the information not captured by the predictors/model.

The regression assumptions are as follows:

- Linearity
    - We assume that each predictor has a linear relationship with the dependent variable. If we don't see a clear linear relationship, we perform variable transformations (in our case, log transformation) so that we observe a linear relationship. Otherwise, polynomial regression may be a possible solution.
- Independence of observations
    - Each observation in the data must not depend on other observations. This way, we can ensure that each observation gives more information about the relationship between the predictors and dependent variable.
- Normality of residuals
    - Residuals should have a normal distribution centered at 0. This is most important for hypothesis tests, point estimation, and confidence intervals for instances with a small sample size.
- Homoscedasticity
    - The variance of error must not depend on any independent variable.
- No multicollinearity
    - Predictor variables should not be highly correlated with one another. Highly collinear variables can make it difficult to understand the effect of each variable separately.

The σ^2 and β_0,..., β_k parameters have to be estimated in multiple regression. We have already talked about the β coefficients above, so here we will explain the variance variable, or σ^2. Calculated by squaring the standard deviation, σ, variance, or σ^2, tells us the spread of errors and is assumed to be normally distributed.

The β parameters are calculated by minimizing the sum of squared errors (SSE), where residuals are the difference between observed and predicted data points. Below is the formula for SSE.

$$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}))^2$$

where ϵ_i is the residual term for the i^{th} observation, $x_1, \ldots, x_k$ and y are the k predictors and dependent variable, respectively, $x_{1i}, \ldots, x_{ki}$ and y_i are the i^{th} observations of $x_1, \ldots, x_k$ and y, respectively, $\hat{y}_i$ is the estimate of the i^{th} observation of y, and β_0 and $\beta_1, \ldots, \beta_k$ are the model intercept and coefficients for predictors $x_1, \ldots, x_k$.

In multiple regression, σ^2 is calculated using the following formula where n is the number of observations and k is the number of predictors. The σ^2 is also known as the mean square errors (MSE).

$$\sigma^2 = \frac{\text{SSE}}{n - (k + 1)} = \text{MSE}$$

The coefficient of multiple determination, R^2, is a ratio that tells us proportion of observed variation in the dependent variable that can be explained by the model (and predictors). R^2 can range from 0 to 1 and a higher value signifies that a model is more successful at explaining the variation in the variable that is being predicted.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ for which n is the number of observations, y_i is the i^{th} observation of y, and $\bar{y}$ is the sample mean of y. SST is the total sum of squares, or the sum of squared deviations about the sample mean of observed y values.

Since the incorporation of multiple predictors will generally increase R^2, we also report the adjusted R^2 that adjusts for the number of predictors, k.

$$R^2_{\text{adj}} = \frac{(n-1)R^2 - k}{n - (k+1)}$$

where n is the number of observations and k is the number of predictors.

In this study, we are testing the hypothesis that the dependent variable, Median House Value (`MEDHVAL`) is independent of all predictor variables, % of residents with at least a bachelor's degree (`PCTBACHMOR`), % of single family houses (`PCTSINGLE`), % of houses that are vacant (`PCTVACANT`), and the number of households living in poverty (`NBELPOV`). We test this hypothesis by using the model utility test, also referred to as the F-ratio or F-test which is a goodness of fit measure. This tests the Null Hypothesis (H_0) that all coefficients in the model are zero, i.e. none of the independent variables is a significant predictor of the dependent variable. The alternative hypothesis (H_a) states that at least one of the predictors has a non-zero coefficient. We can typically reject the null hypothesis when we have a $p < 0.05$ for the F-ratio.

Below is the formula for the F-statistic:

$$\text{F} = \frac{\text{MSR}}{\text{MSE}}$$

where F is the F-statistic, MSR is the mean square regression, defined as $\text{MSR} = \frac{\text{SSR}}{k}$ where SSR is the sum square of regression and k is the number of predictors in the model. $\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ where n is the number of observations, $\hat{y}_i$ is the predicted value of y of the i^{th} observation, and $\bar{y}$ is the sample mean of y where y is the dependent variable.

In addition to the model utility test for the overall model, a hypothesis test is done for each predictor, i, where we test the significance of each beta. These are called t-tests and are usually only examined if the F-ratio is significant. The null hypothesis for these t-tests is that the predictor i is not associated with the dependent variable ($\beta_i = 0$) and the alternative hypothesis is that β_i is non-zero. The null hypothesis can be rejected when $p < 0.05$ here as well, which indicates that that particular predictor is a statistically significant predictor of the dependent variable.

Additional Analyses

Stepwise Regression

In addition to our multiple regression, we will also run a stepwise regression, which is a simple data mining method which selects predictors based on some criteria. In our study, we will use the Akaike Information Criterion (AIC), a measure of relative quality of statistical models to select predictors.

There are a few limitations to stepwise regression. First, the final model is not guaranteed to be optimal, and in fact, there are often several good-quality models, but stepwise regression will only present one such model. Stepwise regression is purely mathematical and does not take into account a researcher's

knowledge about predictors and end up excluding them. There are Type I and Type II errors that can occur with stepwise regression, so we cannot assume that we have kept all the most important variables and gotten rid of all of the unimportant variables. In fact, many t-tests are conducted as part of stepwise regression, which makes the probability of committing a Type I or Type II error high.

K-fold Cross-Validation

We also include a k-fold cross validation in our study (where k = 5). K-fold cross validation is a way to validate our model and examine it's generalizability, or ability to predict data accurately on unseen data. In k-fold cross validation, we first separate the data set into random, k equal-sized groups. The first fold or group is treated as the validation data set, and the model parameters are estimated using data from the other k - 1 folds (which we call the training set). The mean squared error (MSE) is computed for the validation fold and the procedure is repeated k times (one time for each fold), resulting in k estimates of MSE. Then, we average the MSEs across the k folds to calculate the k-fold MSE and take the square root of that measure to compute the k-fold root mean squared error (RMSE).

We use this RMSE to compare it to RMSE of different models to choose the best model (which we define as the model with the smallest RMSE). Below is the formula for RMSE:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}\epsilon_i^2}{n}}$$

where y_i is the i^{th} observation of y, the dependent variable, $\bar{y}$ is the sample mean of y, and n is the number of observations.

Software

We used R for all analyses (R Core Team 2024).

Results

Exploratory Results

Summary Statistics of Variables

Table 1 displays the mean and standard deviation of the dependent variable, median house value; and the independent variables, percent bachelor degree or higher, median household income, percent vacant housing units, percent single family units, and the number of units below poverty level. These summary statistics are calculated from the original data before logarithmic transformation.

Show

Table 1. Summary Statistics of Dependent and Independent Variables

Variable	Mean	SD
Median House Value (MEDHVAL)	$66,287.73	$60,006.08

Variable	Mean	SD
Pct Bachelor or more (PCTBACHMOR)	16.1%	17.8%
Median Household Income (MEDHHINC)	$31,541.76	$16,298.43
Pct Vacant (PCTVACANT)	11.3%	9.6%
Pct Single Unit (PCTSINGLES)	9.2%	13.2%
Units Below Poverty level (NBELPOV100)	190	164

The dependent variable has a mean value of $66,287.73 and a standard deviation of $60,006.08, indicating a broad range of median house values among block groups in Philadelphia. Most of the independent variables have high standard deviations, suggesting that values are spread out from the mean. Our histograms of the distribution of each variable show that most of these variables do not follow a normal distribution, which could explain the high standard deviation. The one exception to this is median household income, which is more normal and has a lower standard deviation.

Variable Histograms

After log-transformation, the dependent variable looks more normally distributed (see figure 1). The same is true for the transformed number of households below the poverty level (see figure 3). The other variables have a large spike at zero after being log-transformed which makes them unfit for our model (see figures 2, 4, and 5). Our regression will use the log-transformed dependent variable `LNMEDHVAL`, and the log-transformed number of households below the poverty level, `LNNBELPOV100.` The remaining independent variables (`PCTBACHMOR`, `MEDHHINC`, `PCTVACANT`, `PCTSINGLES`) will remain untransformed in our regression. Other regression assumptions will be examined in a separate section below Regression Assumption Checks.

Show

Show

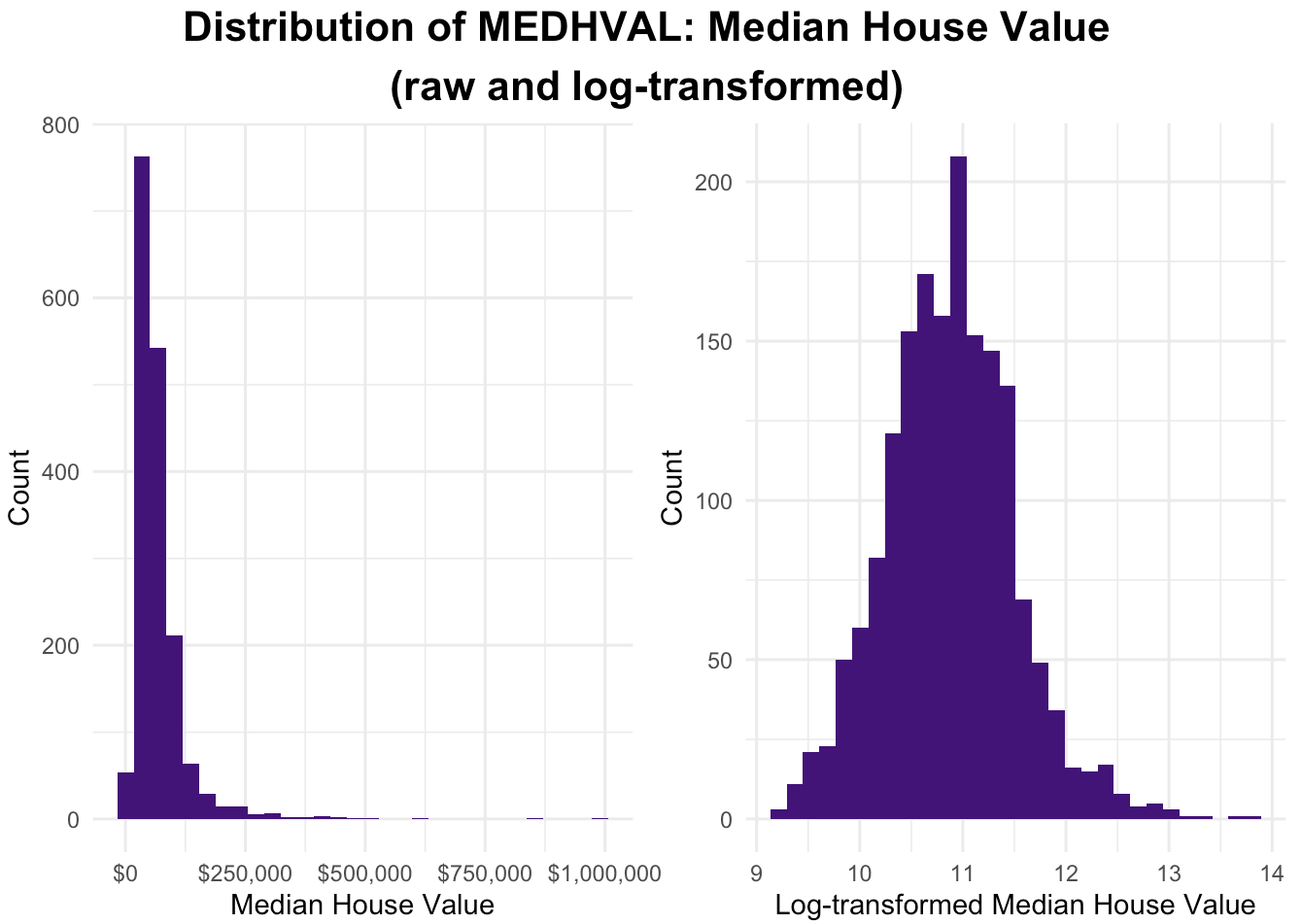

Figure 1.

Show

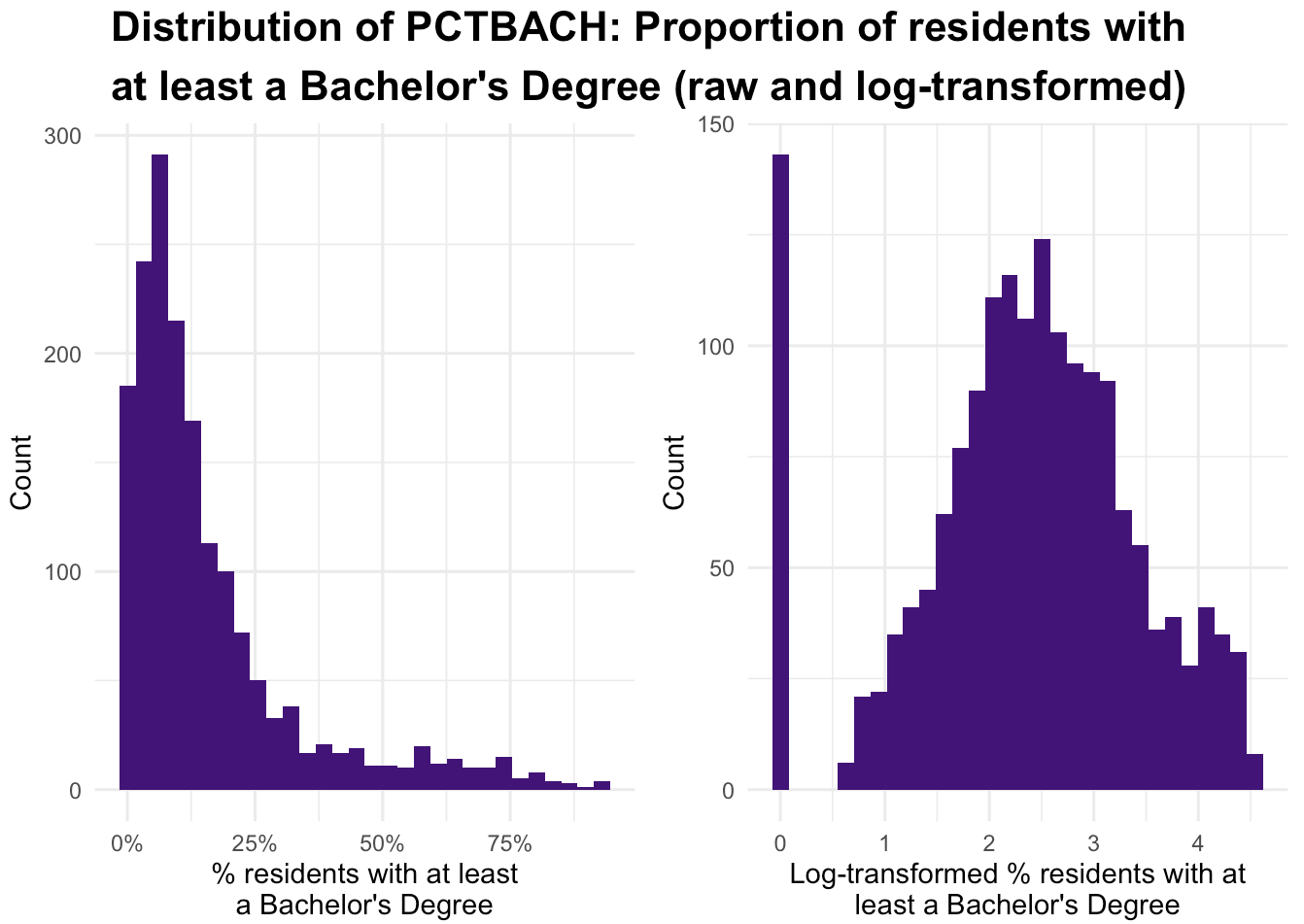

Figure 2.

Show

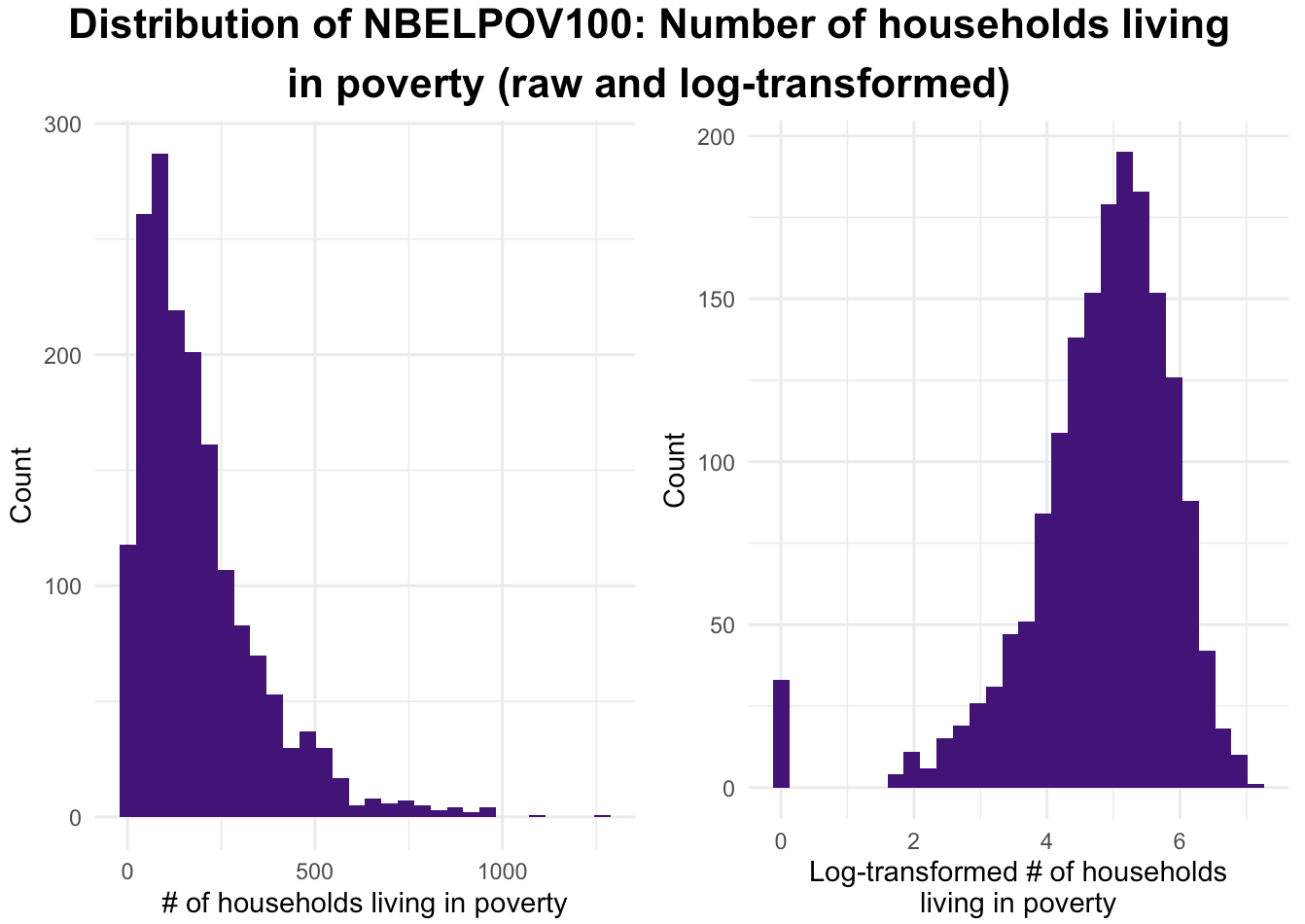

Figure 3.

Show

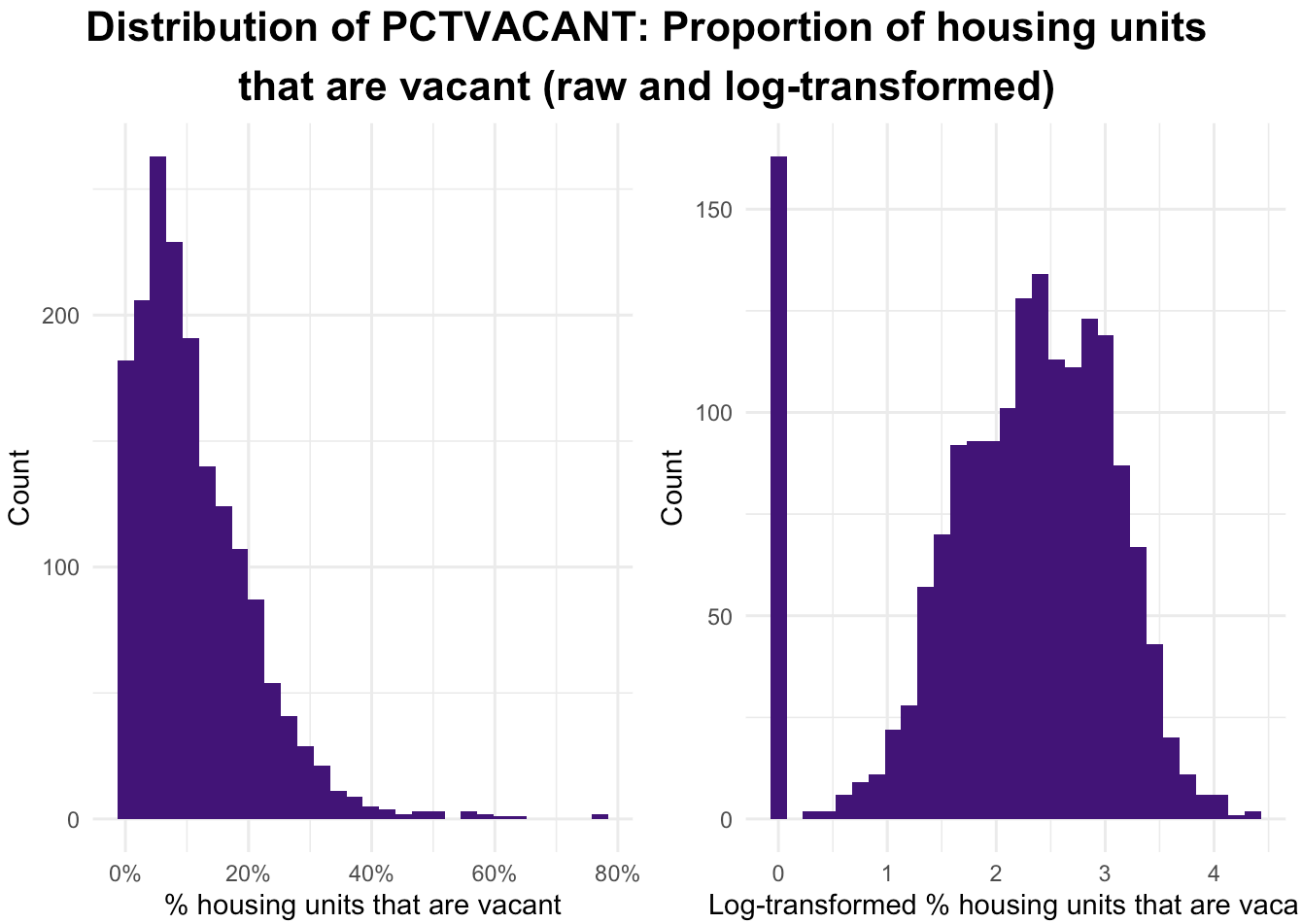

Figure 4.

Show

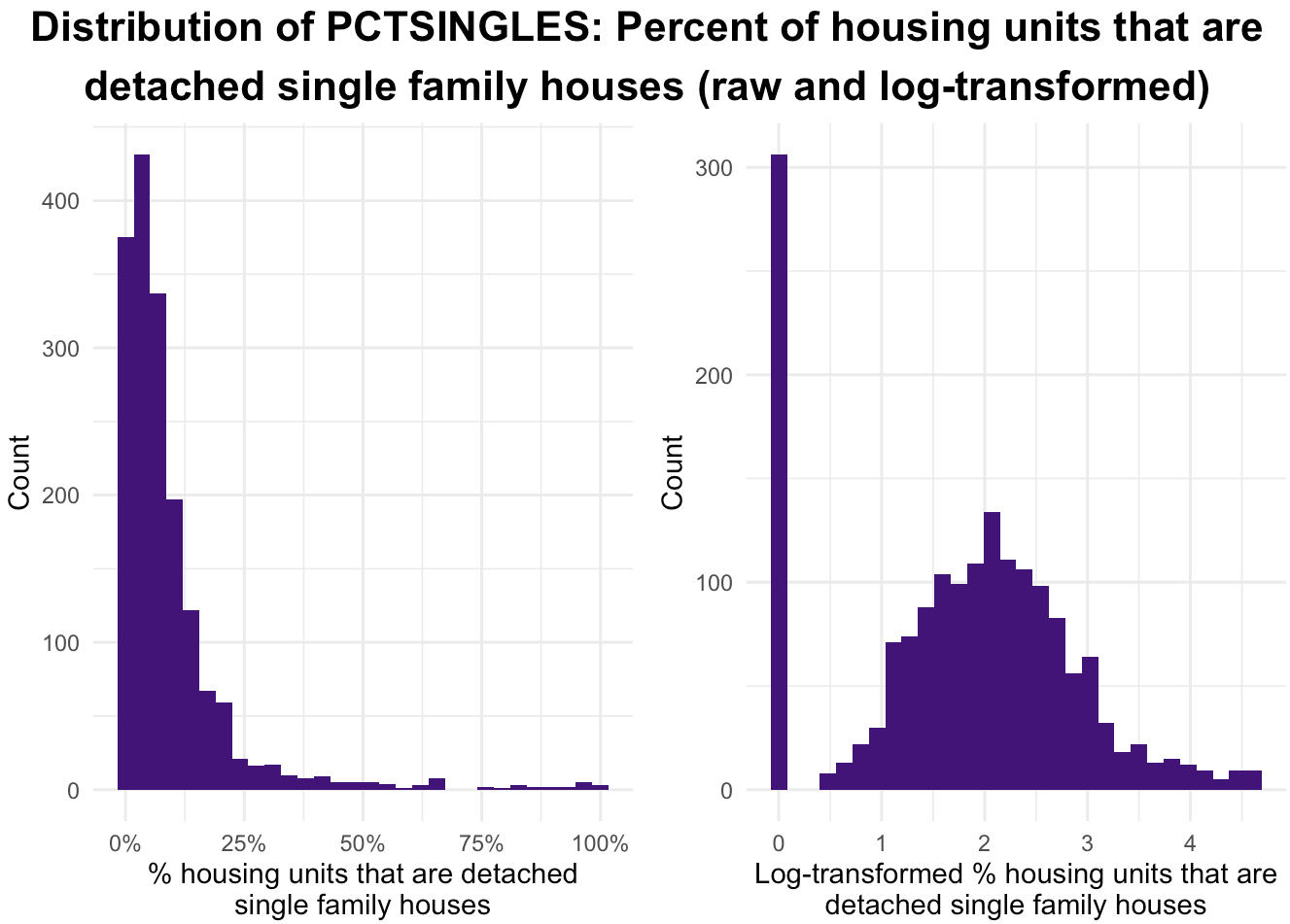

Figure 5.

Chloropleth Maps

The chloropleth map of the dependent variable `LNMEDHVAL` (figure 6) is similar to that of `PCTBACHMOR` with higher values in the northwestern, northeastern, center city, and university city areas. The map of `PCTVACANT` looks like the opposite of `LNMEDHVAL`, suggesting a possible negative relationship between that predictor and the dependent variable. The `LNNBELPOV100` map is a bit hard to interpret, but if we look closely, the highest values of this variable are concentrated in areas similar to the highest values of `PCTVACANT`, again suggesting a negative relationship between this variable and our dependent variable.

Based on the visualizations, none of the predictors appear strongly inter-correlated with one other. The relationships observed from these chloropleth maps were between predictors and the dependent variable, not within the predictors themselves. Though we observe some similarities across the chloropleth maps of predictors, none of them look like copies of each other. Thus, we do not anticipate severe multicollinearity to be an issue in our analysis.

Show

Show

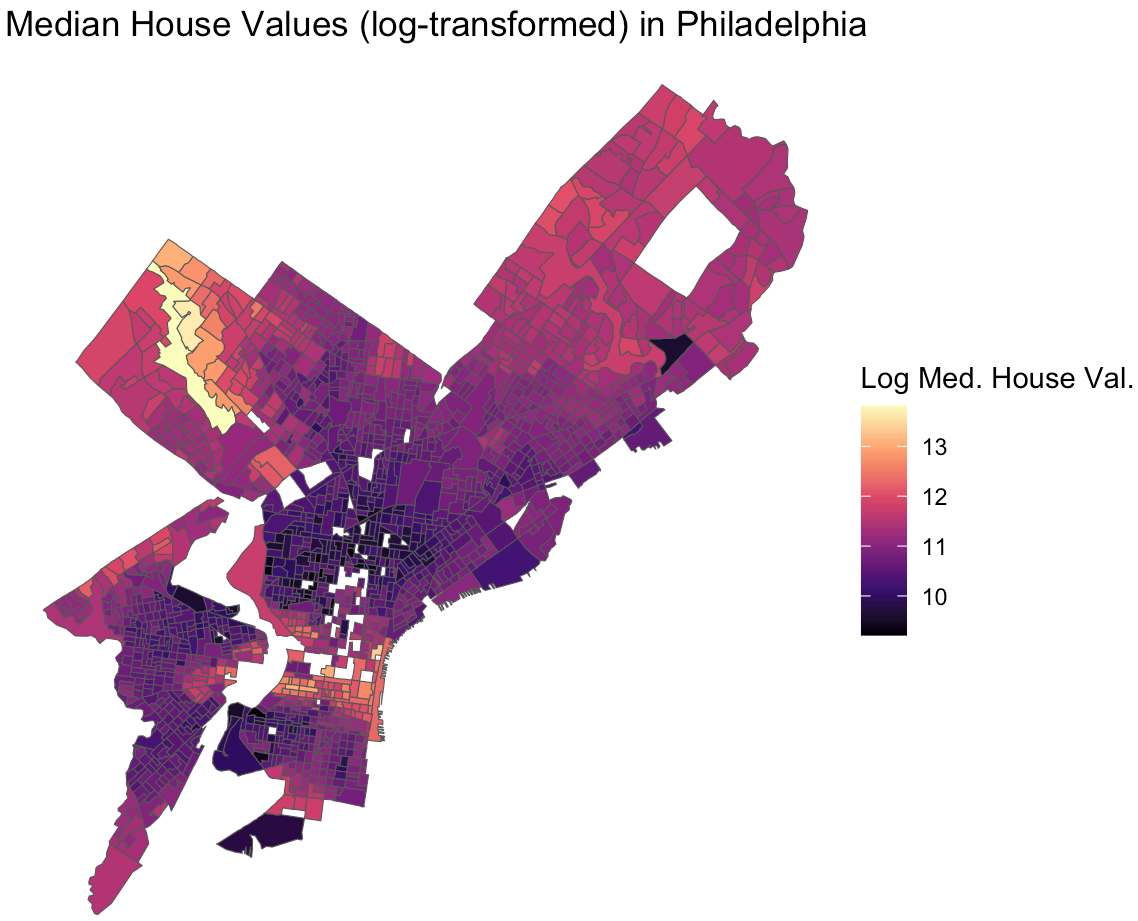

Figure 6.

Show

Chloropleth Maps of Predictors

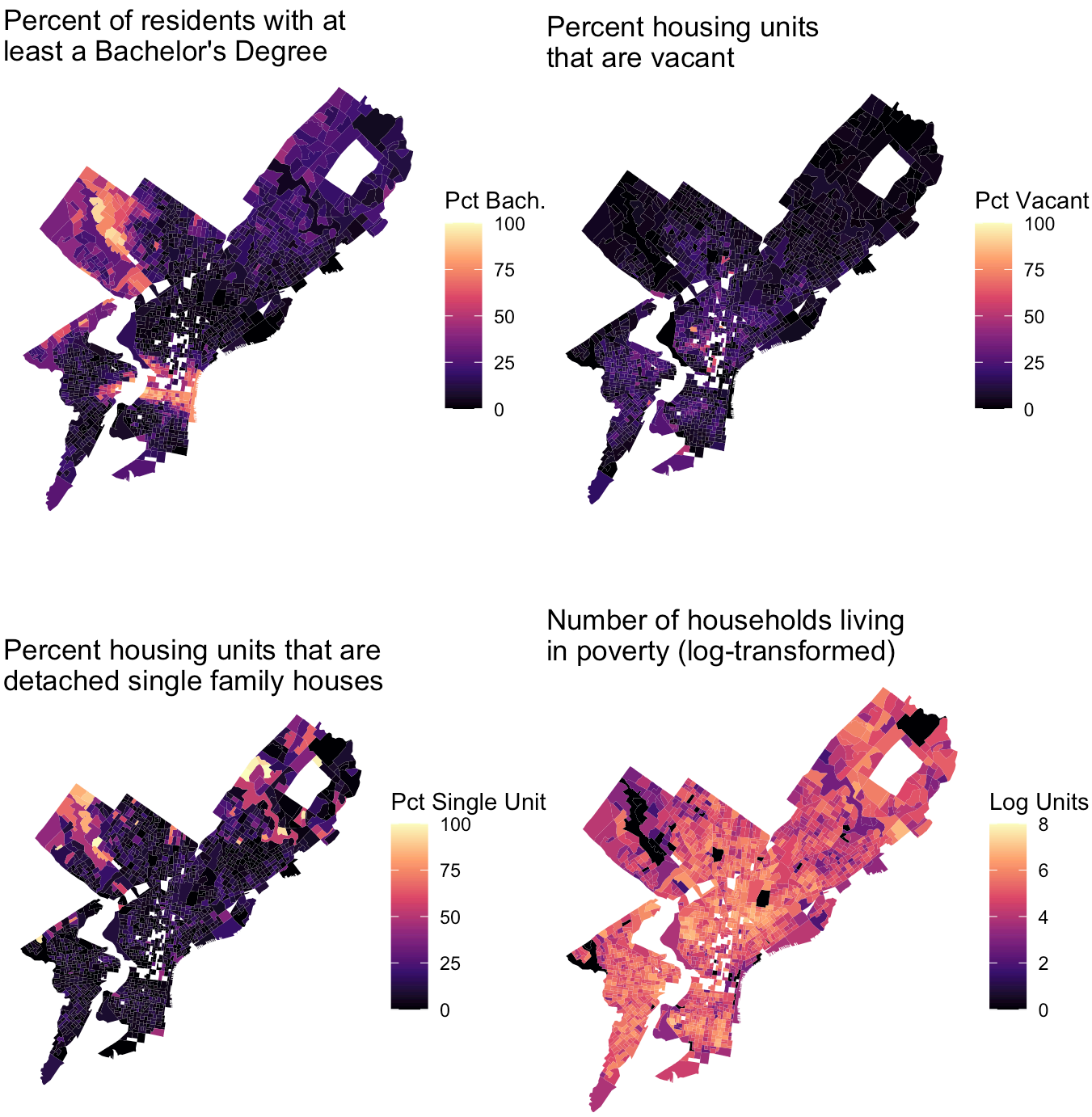

Figure 7.

Correlation Matrix

Show

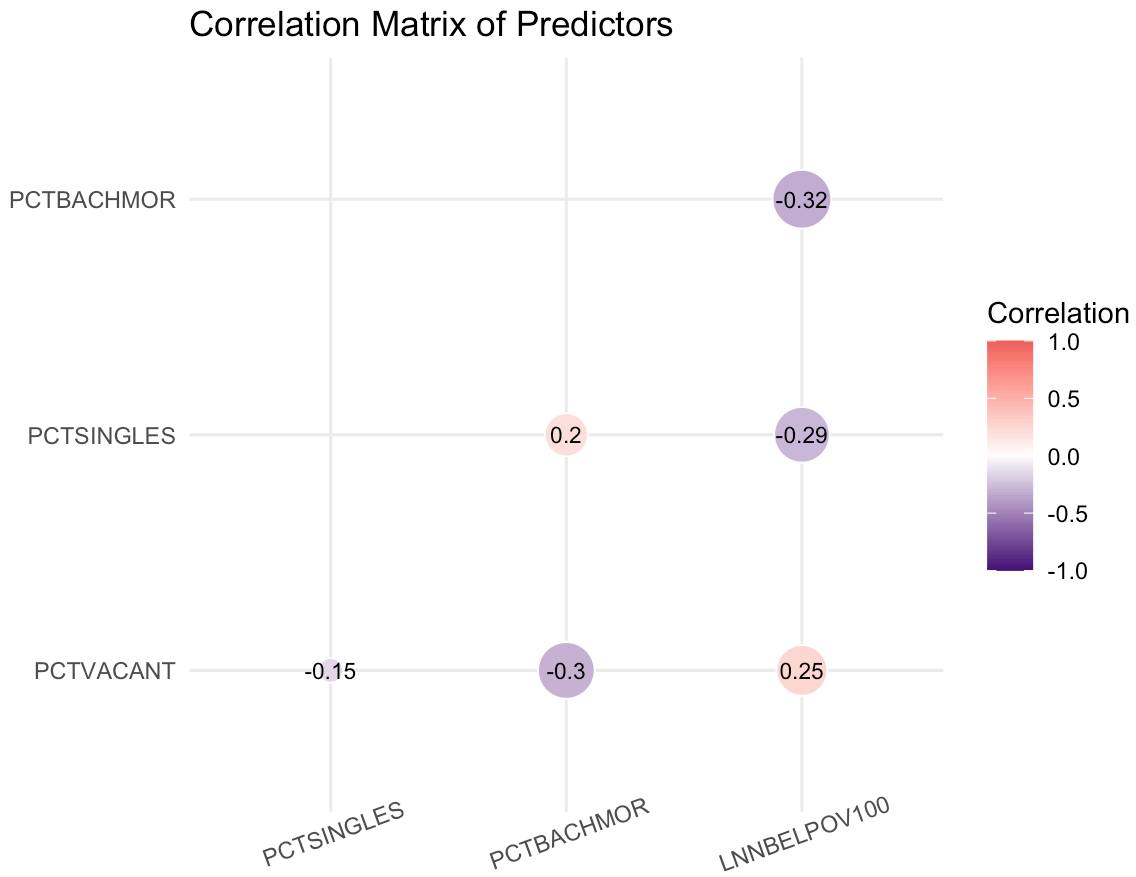

Figure 8.

The correlation matrix above (figure 8) supports the earlier conclusion that there is not severe multicollinearity among our predictors. Correlation values between variables are low, all falling within the absolute value range of 0.15 and 0.32.

Regression Results

Regression Output

We regressed log-transformed median household value (`LNMEDHVAL`) on the % of vacant housing units, % of single family houses (`PCTSINGLES`), % of residents with at least a bachelor's degree (`PCTBACHMOR`), % of vacant housing units (`PCTVACANT`), and the log-transformed number of households living in poverty (`LNNBELPOV100`). Below is the summary of our regression model.

Show

```
##
## Call:
## lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
##     LNNBELPOV100, data = regdata_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25825 -0.20391  0.03822  0.21744  2.24347
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  11.1137661  0.0465330 238.836 < 0.0000000000000002 ***
## PCTVACANT    -0.0191569  0.0009779 -19.590 < 0.0000000000000002 ***
## PCTSINGLES    0.0029769  0.0007032   4.234            0.0000242 ***
## PCTBACHMOR    0.0209098  0.0005432  38.494 < 0.0000000000000002 ***
## LNNBELPOV100 -0.0789054  0.0084569  -9.330 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3665 on 1715 degrees of freedom
## Multiple R-squared:  0.6623, Adjusted R-squared:  0.6615
## F-statistic: 840.9 on 4 and 1715 DF,  p-value: < 0.00000000000000022
```

The quality of the model is quite satisfactory, featuring an R^2 and Adjusted R^2 of 0.662 and 0.661, respectively. This indicates that roughly two-thirds of the variance in median house values can be accounted for by our model. Additionally, the F-test yielded a very low p-value, suggesting that we can reject the null hypothesis that all coefficients in the model are 0.

The regression output tells us that each variable is highly significant ($p < 0.001$ for each variable). `LNNBELPOV100` and `PCTVACANT` are both negatively associated with `LNMEDHVAL`. As the number of units living in poverty (`NBELOV100`) changes by 1%, the expected value of Median House Value (`MEDHVAL`) changes by approximately $(1.01^{\beta} - 1)\cdot 100\% = (1.01^{-0.079} - 1)\cdot 100\% = -0.079\%$, holding all else constant (since $\beta = -0.79$ and $|\beta| < 20$ for `LNNBELPOV100`, we can approximate $(1.01^{\beta} - 1)\cdot 100\% \approx \beta\%$). The coefficient for `PCTVACANT` is -0.019, for which the absolute value is ≤ 0.3, so we can say that as `PCTVACANT` goes up by one unit (in this case 1%), the expected value of `MEDHVAL` changes by approximately -1.9% ($100\beta_1\%$) holding all other variables constant. `PCTSINGLES` and `PCTBACHMOR` are both positively associated with `LNMEDHVAL` with small coefficients, $\beta = 0.003$ and 0.021, respectively (in both cases, $|\beta| \leq 3$), so we can use the same approximation as we did for `PCTVACANT`. As `PCTSINGLES` goes up by one unit with all other variables constant, the expected value of median house value changes by 0.3%. As `PCTBACHMOR` goes up by one unit with all other variables held constant, median house value changes by approximately 2.1%.

Regression Assumption Checks

This section will examine testing model assumptions in our regression. Section 3a Exploratory Results includes the variable distributions for this model.

Scatter Plots

Show

Show

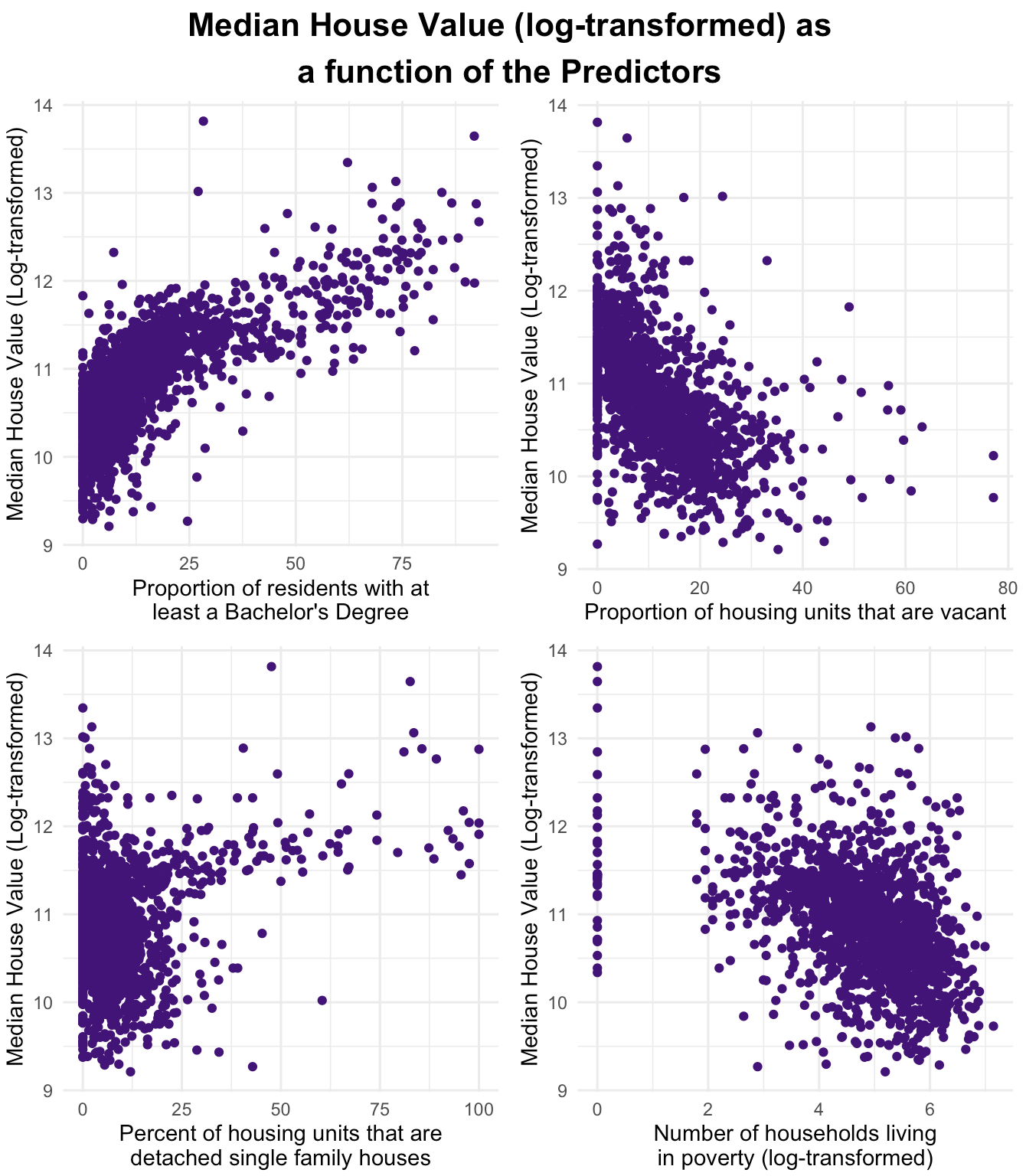

Figure 9.

Figure 9 visualizes the relationship between the dependent variable, the log-transformed median house value (`LNMEDHVAL`), on the y-axis, and each of our predictors on the x-axis. The relationship between house values and the percentage of block residents with at least a bachelor's degree is positive and somewhat linear. The plots for vacant units and the log-transformed number of households in poverty show a negative association with median house values, but it is questionable whether this is a linear relationship,

especially in the case of the plot of log-transformed number of households in poverty where we see several observations with a value of zero for the predictor. The plot of percentage of single family homes is not linear, but a slight positive relationship with the dependent variable can be observed.

Standardized Residuals

We standardize the regression residuals by dividing them by their standard error, making it easier to compare residuals to each other and identify potential outliers. Figure 10 shows that the standardized residuals follow a normal distribution.

Show

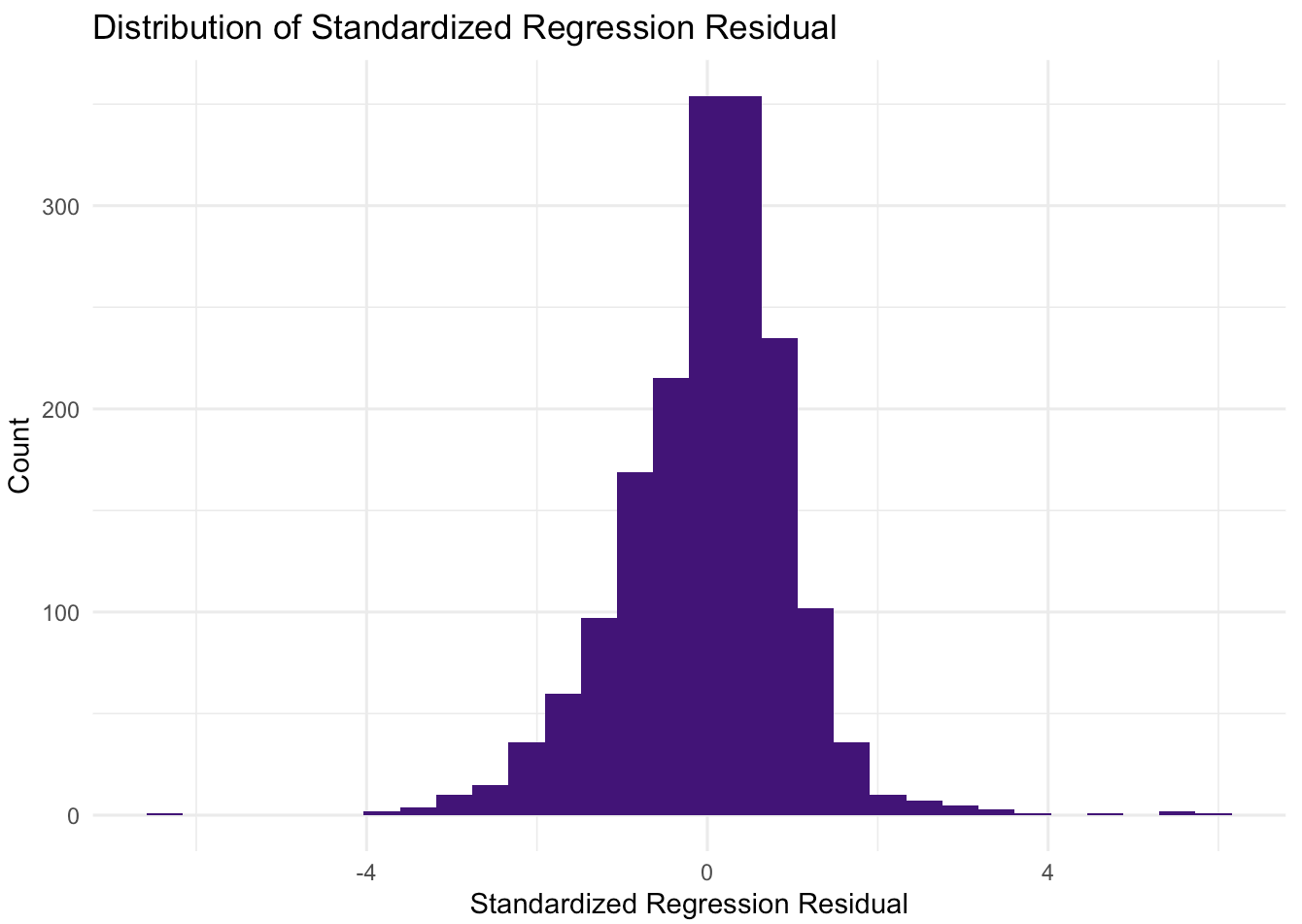

Figure 10.

The scatter plot of standardized residuals by predicted values (figure 11) indicates that there is no heteroscedasticity in our model–that is, there is no relationship or pattern in the scatter plot between predicted values and the standardized residuals. Though the plot appears homoscedastic, we detect the presence of some outliers.

Show

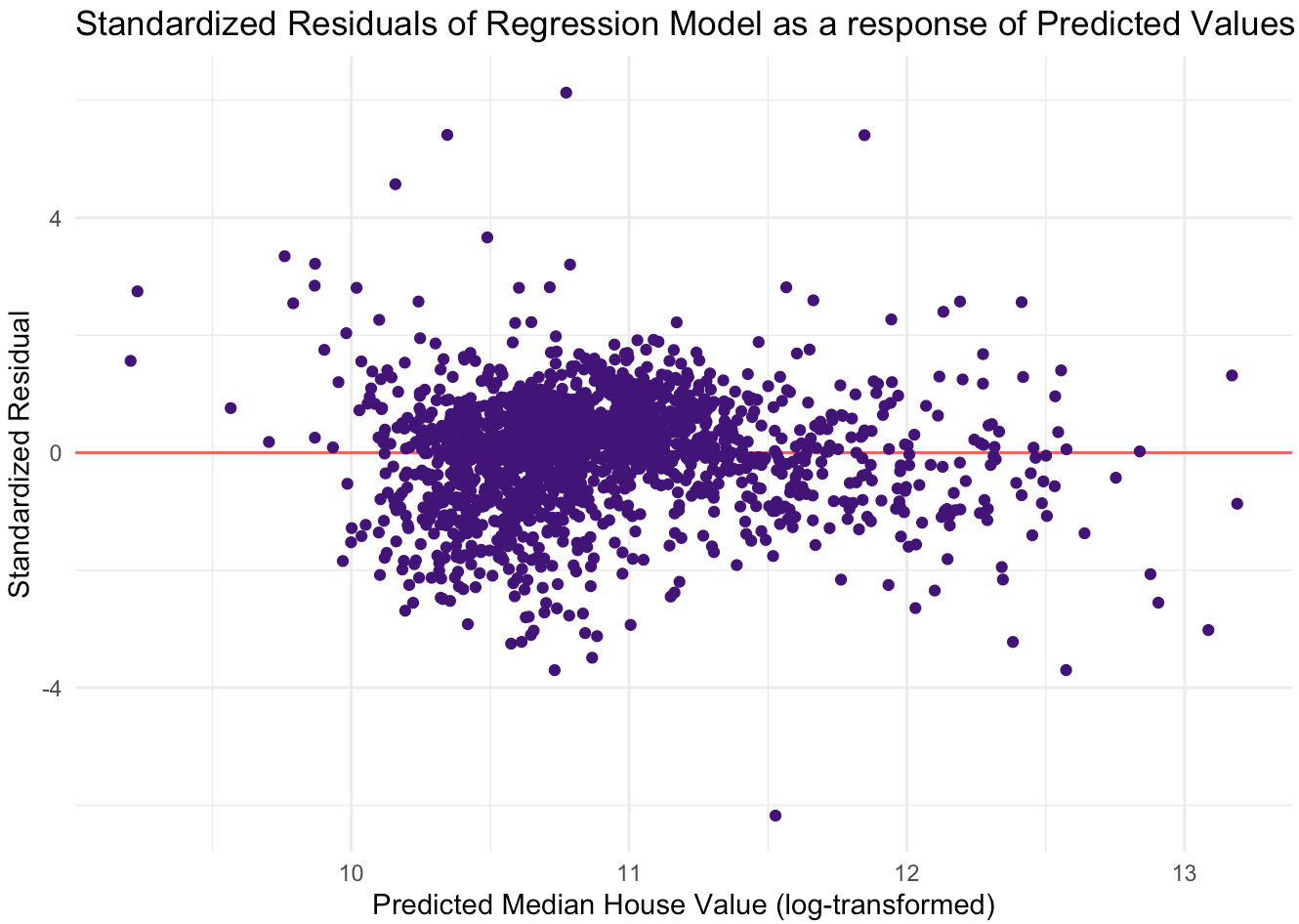

Figure 11.

Show

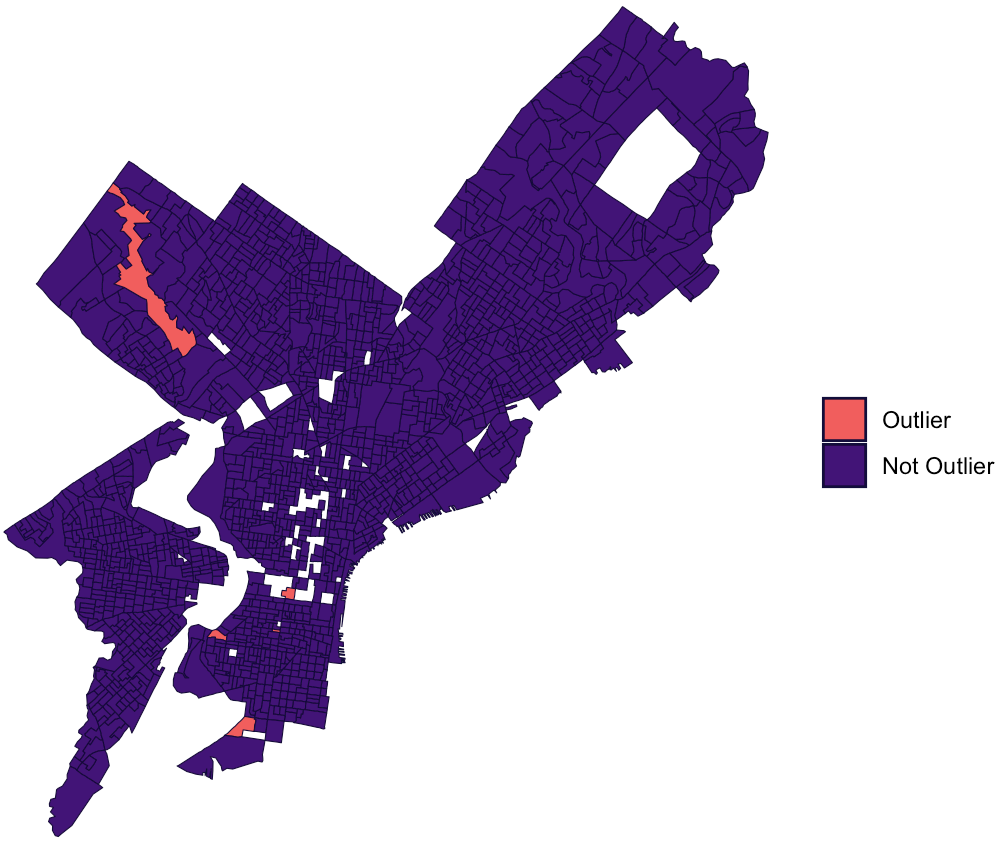

Figure 12.

To identify extreme outliers, we used the absolute value of the standardized residuals and defined an extreme outlier to be an observation where the absolute value of the standardized residual is 4 or greater. A map of these observations (figure 12) reveals that these outliers are located near a large park (Wissahickon Valley park), City Hall, a gas plant, and other non-residential areas. Unique neighborhood characteristics for blocks located near one of these sites could explain the high standardized residuals as they are hard to capture in a general model with the predictors that we chose. This context is useful for understanding our model, but does not prompt any additional action.

The maps of the dependent variables and predictors did show spatial autocorrelation. The map of the dependent variable showed blocks with high home values next to other blocks with high home values, and vice versa. Maps of independent variables indicated that blocks with a high percentage of people with bachelor's degrees were close to each other, and that blocks with high poverty are not usually near blocks with little poverty. Overall, the chloropleth maps show that block groups with similar characteristics tend to cluster together in space.

Show

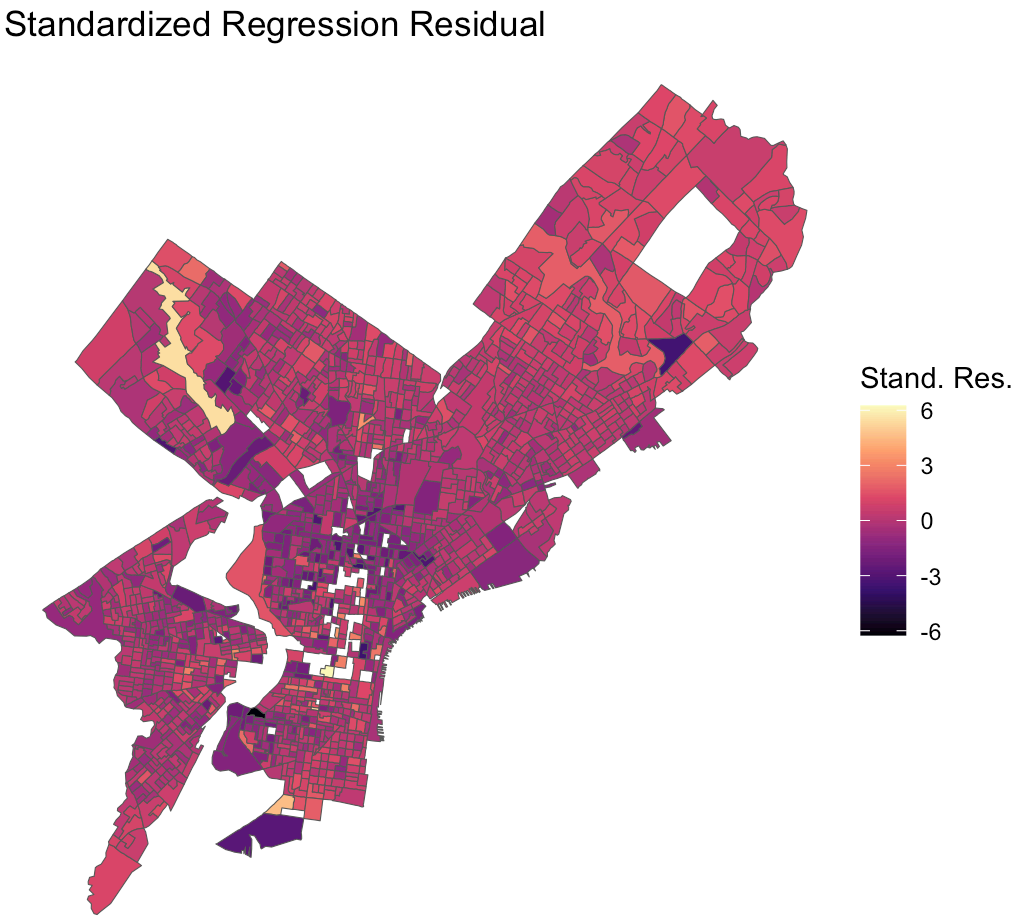

Figure 13

This chloropleth map of the standardized residuals also has a pattern of spatial autocorrelation. Areas with negative standardized residuals tend to be located near each other, as is true for areas with positive standardized residuals. Our standardized residuals follow a normal distribution and there is no heteroscedasticity, but the values do cluster together in space.

Stepwise Regression

Show

```
## Start:  AIC=-3448.07
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100
##
##                Df Sum of Sq    RSS     AIC
## <none>                      230.34 -3448.1
## - PCTSINGLES    1     2.407 232.75 -3432.2
## - LNNBELPOV100  1    11.692 242.04 -3364.9
## - PCTVACANT     1    51.546 281.89 -3102.7
## - PCTBACHMOR    1   199.020 429.36 -2379.0
```

Show

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100
##
## Final Model:
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100
##
##
##   Step Df Deviance Resid. Df Resid. Dev       AIC
## 1                       1715   230.3435 -3448.073
```

All 4 predictors are kept in the final model based on the stepwise regression results, because the model with all predictor variables included has the lowest AIC (Akaike Information Criterion) value and lowest RSS (residual sum of squares) value. The AIC is essentially an estimator for the average error on a set of predictions, whereas the RSS is the actual sample error of predictions. For both, we want the values to be as low as possible.

Cross-validation

Show

Table 2. Comparing RMSE of Original and Simplified Models

	Original Model	Simplified Model
Root Mean Squared Error (RMSE)	0.36752	0.44382

"Original Model" refers to the model that uses all four predictors (PCTVACANT, PCTSINGLES, PCTBACHMOR, and LNNBELPOV100) to predict LNMEDHVAL, while "Simplified Model" refers to the model with only two predictors: PCTVACANT and MEDHHINCOME.

Table 2 shows the results of the k-fold cross validation (using k = 5 folds) for our main regression model and simplified model suing only `PCTVACANT` and `MEDHHINCOME` as predictors for `LNMEDHVAL`. Our original model performed better than the simplified one, as it has a lower RMSE.

Discussion and Limitations

Conclusions

In summary, we investigated the predictive power of four variables (percentage of vacant homes, percentage of people with bachelor's degrees, number of households below poverty, and percentage of single-family detached homes) when the area median house value was regressed on them using a linear model (using 2000 census data in the city of Philadelphia, Pennsylvania). To ensure that the assumptions of the linear model were being met, we explored the data prior to running a regression model. We investigated the distributions of our variables and decided from there to use a log-transformed version of median house

value and number of households living in poverty in order to preserve normality. We also checked that there was no correlation between predictor variables, although we saw that each of these variables had spatial correlations; they had a relationship with spatial neighborhoods.

We then regressed median house value on all four predictors and observed a high R squared value of around 0.66, meaning that about two thirds of the variation in median house value was explained by our model.The model results showed that all four variables were highly significant to predict median house value, and this was enforced by the results of our stepwise regression that showed the lowest error rate for the model with all 4 variables. This model also outcompeted a model where median house value was regressed only on percent vacancy and median household income, which we checked by comparing RMSE from k-folds cross validation where k = 5. Lastly, we saw that our model residuals were normally distributed and homoscedastic with an average close to zero, which holds mode assumptions. The residuals are not spatially independent, however.

In the stepwise model, both AIC (Akaike Information Criterion) and RSS (residual sum of squares) values are used to compare models with different sets of predictors. The AIC serves as an estimate of the average error across a set of predictions, while the RSS reflects the actual sample error of those predictions. In both cases, lower values are preferred. The stepwise regression results show that the model with all four original predictor variables kept yields the lowest AIC and RSS, which shows that the model is of higher quality than one with redundant variables that would have shown to hinder a low AIC or RSS.

The RMSE was about 0.37 for the original 4-predictor model and about 0.44 for the model where median house value was regressed on just median household income and percent vacancy. This shows that there was greater error in the predictions of the model with only two predictor variables.

Limitations of Model

The model has some limitations. As stated before in the exploratory analysis, all of the variables did not exhibit normality. While we were able to correct this for median house value and number of households in poverty with a log transformation, it makes interpretation of the beta coefficients less easy, with additional calculations required to understand the effect of a variable in an intuitive way. Three of the predictor variables had a spike at 0 when log-transformed, so their original form was used in the model, but that does not violate any assumption for the model so it is fine.

Another limitation to the model is that the residuals are not spatially independent. This shows that the model predicts certain neighborhoods more accurately than others. A potential solution to this would be to add a spatial variable to the model.

Number of households in poverty (`NBELPOV100`) is a variable with raw numbers rather than percentages. This is not scalable to the other predictor variables, which are all in percentages. Furthermore, it is highly influenced by the population in a block group. For example, for a block group with a low percentage of households in poverty, it may be the case that the block group is larger in population than other groups and therefore appear to have “more” households living in poverty in one area than another, when in reality they may be very sparse. This may cause interpretation errors.

Ridge or Lasso?

Ridge or LASSO regression do not make sense in this study, because they are tailored to regress a variable on a large number of predictor variables, particularly when the number of observations is not high compared to the number of variables. That is not the case in our study. We have only 4 predictors, and

1,720 observations. In addition, all 4 of our predictor variables appear very significant to the prediction of median house value, so we would not want any of them weighted lower, with a beta coefficient approaching zero in ridge regression or being set to zero in lasso regression.

References

Bureau, U. S. Census. 2000. “American Community Survey 2000 Decennial Data.”

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/).