Using OLS Regression to Predict Median House Values in Philadelphia

Shengao Yi 76852392, Zhonghua Yang 74788569, Xuening Zhang 65401665

1 Introduction

Based on Philadelphia data in Census block group level, this report will assess the relationship between the median house values and several neighborhood characteristics, and predict the median house values in Philadelphia (PA) using Ordinary Least Square (OLS) regression with the help of R studio software. The house value is a parameter that could represent the interests of people to estate in different fields. The spatial analysis of this project is based on a dataset of census data with various variables at the census block group level, including 1816 samples with 7 features, such as census block group id, median value of all owner occupied housing units, and median household income. The multiple OLS regression and step wise regression is applied to examine the relationship between the house price and characteristics including education, vacancies, house forms, and poverty. Based on the spatial analysis, maps are created in R studio.

The existing and projected house-owners, and private sectors are apparently interested on the understanding of which and how neighborhood characteristic affect the house values. As well as that, both city governments and justice-oriented are interested in the elements that influence house values. In this project, the effect of variables including education, poverty, vacancies, household composition on the price of houses in Philadelphia. As is known to all, the house prices are relatively low in the areas with a high poverty level. Also, the paper of Painter and Yu (2008) states that the higher house values are associated with the family with a higher level of education, considering the

exorbitant cost of higher education in the United States. Additionally, in a perfect competition housing market, house prices are determined by supply and demand. If the demand for dwellings exceeds the supply of dwellings, there will be a housing shortage and house prices will rise. The vacancy rate will be low in that case because house hunters will quickly occupy vacant dwellings (Hoekstra and Vakili-Zad, 2011). Furthermore, the house price is likely to be higher in the areas with more detached single family houses. In this report, these variables are examined to obtain the relationship between the house values and the neighborhood variables statistically.

2 Methods

2.1 Data Cleaning

This project uses census data as the basic spatial analysis unit, which includes various demographic variables at the census block group level, containing 1816 samples in the original dataset with 7 features, such as census block group id, median value of all owner occupied housing units, median household income and so on. Then, to make the model prediction results more accurate, we filtered out block groups with small populations (<40), no housing units, the median house value less than $10,000. In addition, two separate block groups in North Philadelphia were removed due to a very high median house value (greater than $80,000) and very low median household income (less than $8,000). At last, there were 1720 groups left can be used in this analysis.

2.2 Exploratory Data Analysis

Before we get through the analysis, we should give a glimpse to the data and calculate summary statistics. The statistics include the mean and standard deviation of both our dependent variable Median House Value (MEDHVAL) and four independent predictors:

- **PCBACHMORE:** Proportion of residents in Block Group with at least a bachelor's degree.
- **PCTVACANT:** Proportion of housing units that are vacant.

- **PCTSINGLES:** Percentage of housing units that are detached single family houses.
- **NBELPOV100:** Number of households with incomes below 100% poverty level (i.e., number of households living in poverty).

The statistics can help us understand what the values our predictors and dependent variable look like. Additionally, we will explore the distribution of our data through histograms, which can find out whether they are normally distributed. Detecting the normal distribution of data is important. There are some variables that were substantially skewed to the left or right, which would lead to heteroscedasticity in the regression analysis (Kaushal and Shankar, 2021). At that time, we may need to apply the logarithmic transformation to other variables whose skewness is greater than 1.

Correlation is a standardized measure that informs us how strong the linear relationship is between two variables. It's usually shown as r, which ranges from -1 to 1. When $r > 0$, means positive relationship, $r < 0$ means negative relationship. A value of 0 implies there is no linear relationship at all.

The Pearson's Correlation Equation 1 is as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (1)$$

In which, x_i, y_i are the ith value of x, y, $\bar{x}, \bar{y}$ are the mean value of x, y, r is the correlation coefficient. Typically, if any two predictors have a correlation of $-0.8<r>0.8$, then we should remove them from the model due to the multicollinearity, which is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

2.3 Multiple Regression Analysis

This project will use multiple Ordinary Least Squares (OLS) regression to examine the relationship between our dependent variable and a set of explanatory variables. It's a type of linear least squares method for choosing the unknown parameters in a linear regression model by the principle of least squares. We will calculate the coefficient β_i of each predictor, which can be interpreted as the amount by which the dependent variable y (Median House Value), changes as the independent variable increases by one unit, holding all other predictors constant. The term ε is usually referred to as the residual or random error term in the model, which is defined for each observation i, as the vertical distance from the observed value and the predicted value of y. We are regressing median house value on number of households living in poverty, proportion of residents with at least a bachelor's degree, percentage of individuals, percentage of vacant homes, and percentage of single house units in Philadelphia. Our regression equation 2 is as follows:

$$y = \beta_0 + \beta_1 LNBELPOV100 + \beta_2 PCTBACHMOR + \beta_3 PCTSINGLES + \beta_4 PCTVACANT + \varepsilon \quad (2)$$

OLS regression models have a lot of assumptions.

1. **Linear relationship:** There exists a linear relationship between each dependent and explanatory variable.
2. **Independence:** The observations are independently and identically distributed. Also, the error term ε is independent of the explanatory variables. They are uncorrelated.
3. **Multivariate Normality:** The error term ε is normally distributed conditional on the explanatory variables, which is important for point estimation.
4. **Homoscedasticity:** The error term ε is homoscedastic, which means the residuals have constant variance at every point in the linear model.
5. **No Multicollinearity:** OLS regression has no multi-collinearity, which means that there should be no strong linear relationship (r>|0.8|) between the

independent variables. If the relationship (correlation) between independent variables is strong (but not exactly perfect), it still causes problems in OLS estimators and weakens the model.

The parameters of multiple regression are coefficients β_0, …, β_k, where k is the number of predictors. β_0 means the y-intercept of the regression line. OLS model also calculates the sum of least squares of residuals. Least squares estimates for $\hat{\beta}_0$, …, $\hat{\beta}_k$ are obtained when the quantity of SSE (equation 3 below) is minimized.

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki} \right) \right]^2 \quad (3)$$

The parameter σ^2 determines the amount of variability inherent in the regression model, which is need to be estimated, calculated as equation (4):

$$\sigma^2 = \frac{SSE}{n - (k+1)} = MSE \quad (4)$$

where k is number of predictors, n is number of observations, and MSE stands Mean Squared Error. Also in multiple regression, $SST = \sum (y_i - \bar{y}_i)^2$, $R^2 = 1 - \frac{SSE}{SST}$. Here, R^2 is the coefficient of multiple determination, or the proportion of variance in the model explained by all k predictors. The R^2 will increase with more predictors in the model and can be adjusted for the number of predictors with the equation 5:

$$R_{adj}{}^2 = \frac{(n-1)R^2 - k}{n - (k+1)} \quad (5)$$

Then, to determine a goodness of fit measure, we conduct a so-called model utility test, referred to as the F-ratio, or the F-test. It can be interpreted as a significance test of R-squared. F-ratio tests the Null Hypothesis, H_0 that all coefficients in the model are (jointly) zero vs. Alternative Hypothesis H_a that at least 1 of the coefficients is not 0. In other words, the test is to make sure that we can reject the H_0 for H_a, it occurs in general when $p < 0.05$. The p-value is the probability of observing a value that is at least as different from 0 (the value stated in H_0) as the given estimated value.

After the F-test., we will run a T-test for each individual predictor. For this project, the Null Hypothesis H_0 states that the predictor has no relationship with the dependent variable and our goal is to reject it, support the alternative hypothesis H_a, that is, $\beta_i \neq 0$ for each predictor.

2.4 Additional Analyses

One of mostly used methods is the stepwise regression, which is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model based on some criteria: P-value below a threshold; Smallest value of the Akaike Information Criterion (AIC) and so on. Stepwise regression also has various limitations. The principal drawbacks of stepwise multiple regression include bias in parameter estimation, inconsistencies among model selection algorithms, an inherent (but often overlooked) problem of multiple hypothesis testing, and an inappropriate focus or reliance on a single best model (WHITTINGHAM et al., 2006).

K-fold cross validation is another method we used in this project. It a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. The general procedure is as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k groups.
3. For each unique group:
    a) Take the group as a hold out or test data set.
    b) Take the remaining groups as a training data set.

c) Fit a model on the training set and evaluate it on the test set, calculate the Mean Square Error (MSE) in equation 6:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (6)$$

d) Retain the evaluation score and discard the model.

4. The k-fold MSE estimate is computed by averaging the MSEs across the *k* folds.

$$Test\ MSE = \frac{1}{k}\sum_{i=1}^{k} MSE_i \quad (7)$$

The biggest drawback of k-fold cross validation is the scope of over fitting the model and will also need higher training time if the dataset is too large. This procedure results in k estimates of the MSE, then we can get a statistic known as the Root Mean Squared Error (RMSE), which is an estimate of the magnitude of a typical residual.

In this project, we set K to 5 and compare the RMSE values for different models and choose the model with the smallest RMSE as the best.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{n}} \quad (8)$$

In which, y_i is the i[th] observed value of dependent variable y, $\hat{y}_i$ is the i[th] predicted value of dependent variable. ε is residuals, n is the number of observations in each fold.

We use the software R studio for the data analysis, and creating maps.

3 Results

3.1 Exploratory Results

Firstly, to gain a better understanding of our observations, we have calculated the mean and standard deviation values of the dependent variables (Median House Value) and four predictors: the percentage of people with bachelor's degrees or higher (PCBACHMORE); the proportion of vacant homes (PCTVANCANT); and the

proportion of single-family homes (NBELPOV100) that are poor (PCTSINGLES).

Variable	Mean	SD
Median House Value	66287.73	60006.08
# Households Living in Poverty	189.7709	164.3185
% Individuals with Bachelor's Degrees or Higher	16.08137	17.76956
% Vacant Houses	11.28853	9.628472
% Single House Units	9.226473	13.24925

Table 1. The statistics of the original data.

According to Table 1, it is suggested that the standard normal distribution of the mean is zero and the standard deviation is 1, which is slightly larger, all data of the standard deviation are close to or greater than the average of the variables.

The original data distribution histograms are shown in Figure 1, from which we can find that all of these variables are not normally distributed, the linear regression model assumes that for the residuals of observations in our sample, each variable is normally distributed. If the residuals of our variables are not normally distributed, then we would be violating the assumption. So we need to make the variables normal through Log Transformations.

We use the natural log (or natural log +1, when there are zeros in the observations) for each variables. The natural logarithmic distribution of these variables' histograms are

shown in Figure 2, it can be found that the logarithmic transformation only helps normalize the NBELPOV100 and MEDHVAL variable, since the other predictors have a large spike at zero after transformation. So we will use the original data for other variables in the regression. We will make more regression assumption checks further in this project.

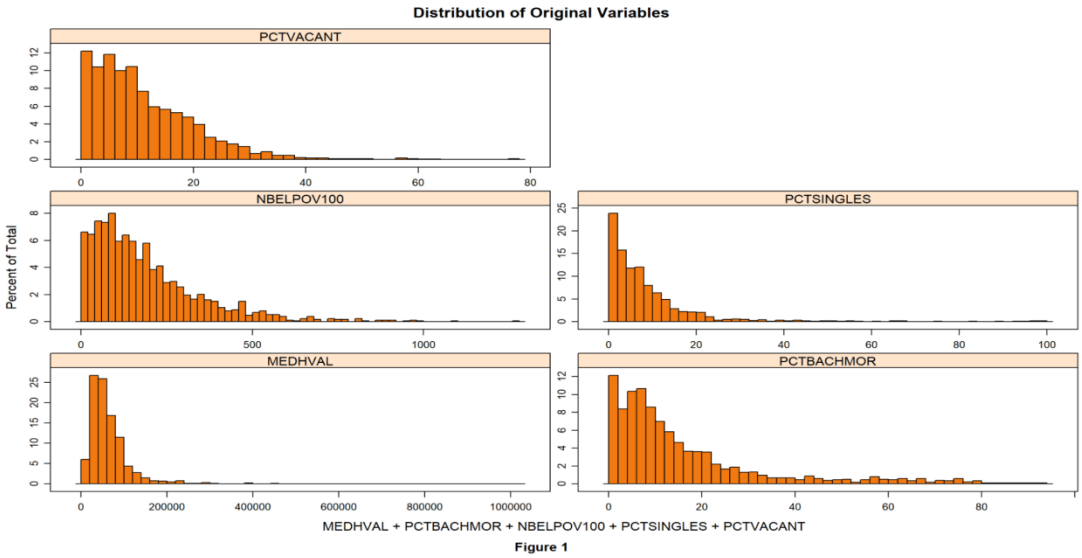


Figure 1. The histograms of the distribution of the original variables, including PCTVACANT, NBELPOV100, PCTSINGLES, PCTBACHMOR, and MEDHVAL.

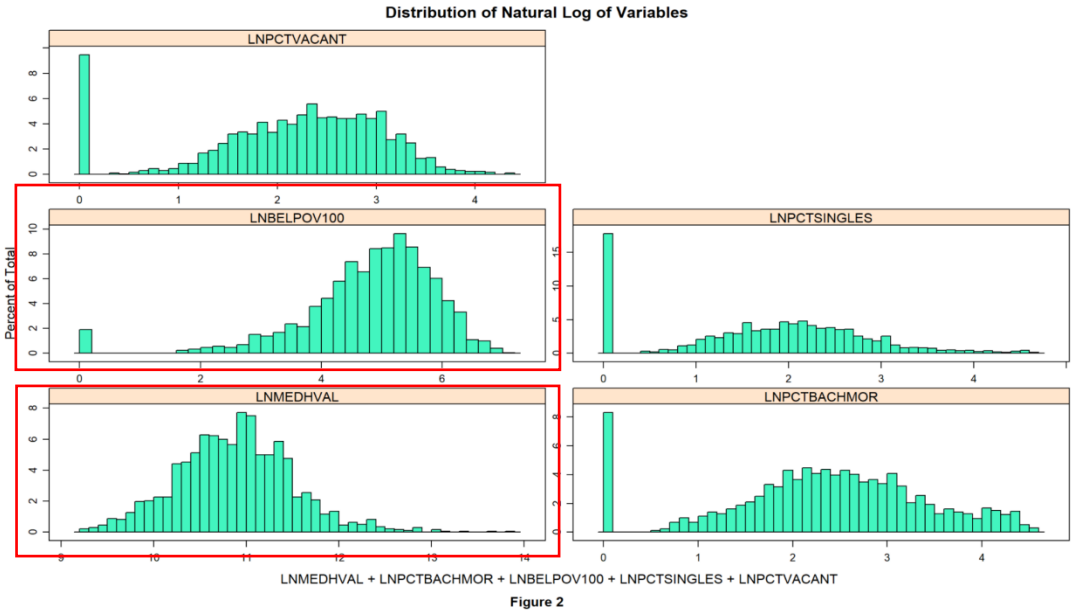
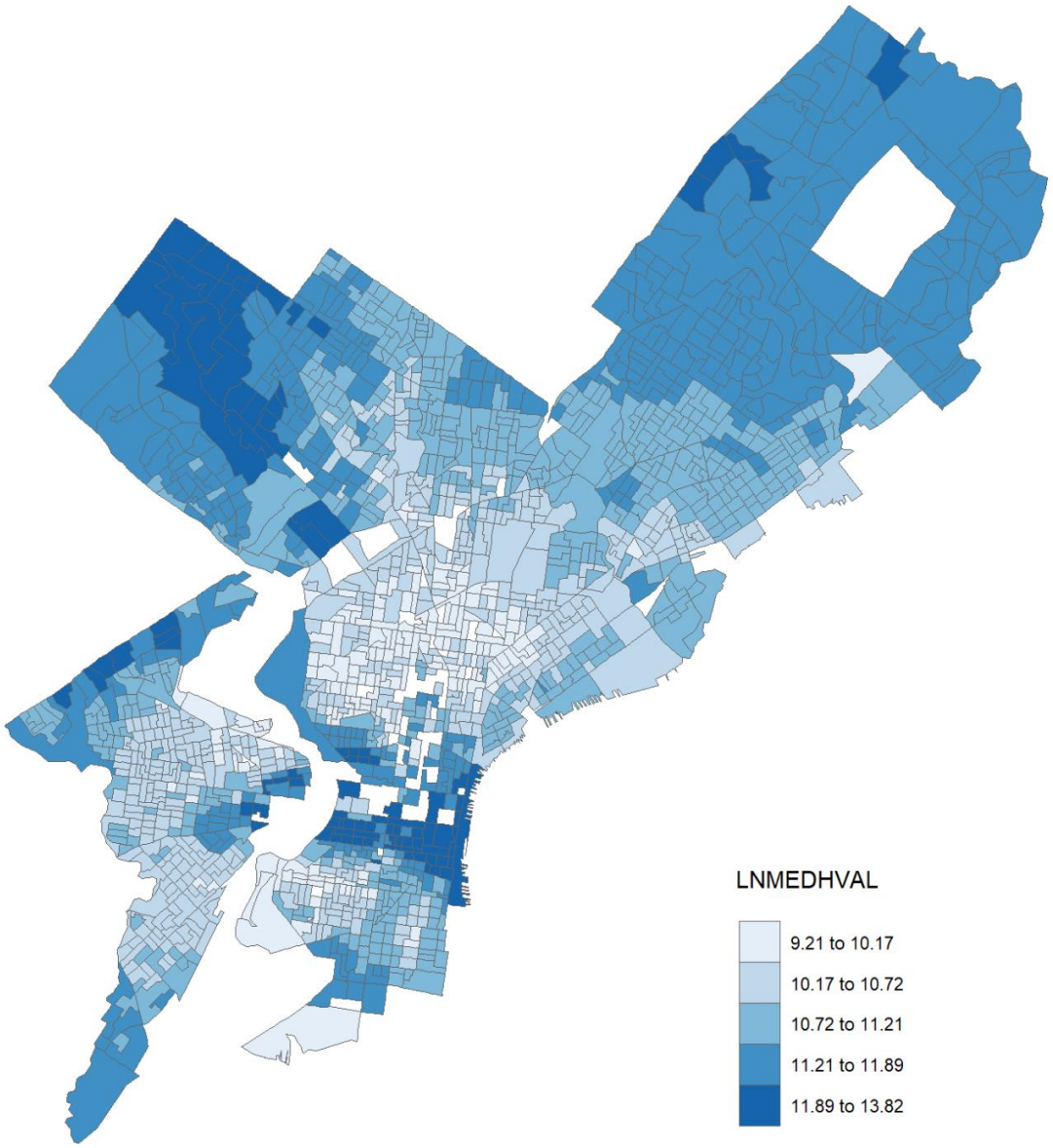
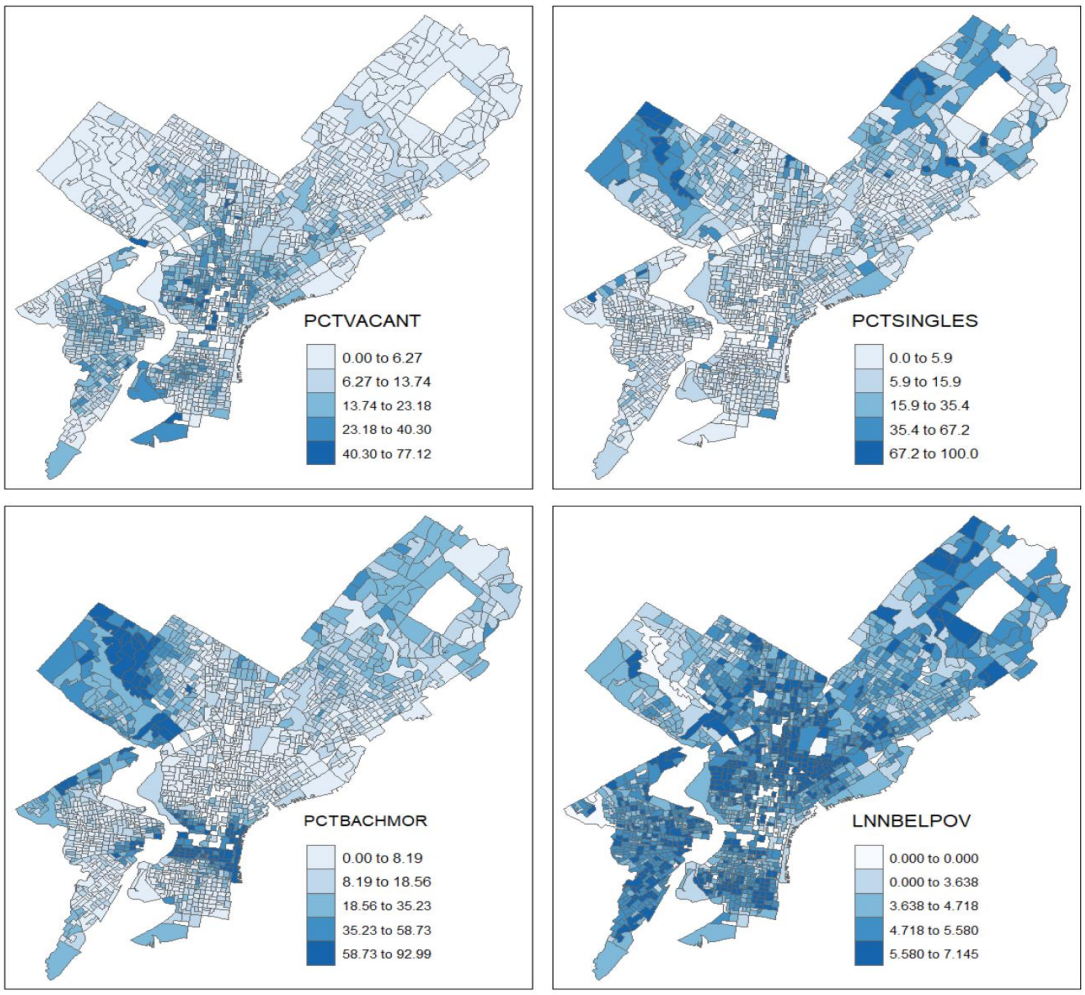


Figure 2. Distribution of natural log of variables. (Only MEDHVAL and NBELPOV100 are transformed in the regression)



Map 1. Different Median House Values across each census area.



Map 2. The predictor variables across different regions.

Now we look into the geospatial characteristics and relationships between the dependent variable and the independent variable from the choropleth maps. It appears that the geographic distribution of vacant properties and the distribution of the median house value are opposite, indicating that there appear to be fewer vacant properties in areas with lower house prices and more vacant properties in areas with higher house prices. That makes sense logically. Additionally, the plot for the percentage of people with a bachelor's degree or higher resembles our dependent variable the most, while the plot for households with low incomes almost exactly resembles the opposite. It means that the high educated people will be more likely to own an expensive house. The prediction ability of the model is confined to elucidating the relationship between variables. It’s unable to substantiate that there is a mathematical relationship between

them. The next step is to assess whether our model assumptions will be affected by the relationship between our predictor variables, such as our poverty and univariate variables.

The correlation coefficients between the predictor variables are listed in Table 2. The correlation coefficient between each variable and itself is found to be 1, which is consistent with mathematical reasoning. In addition, conclusions should be drawn from these data using the correlation coefficients between the predictor variables. Given that multicollinearity should be avoided as much as possible in a multivariable regression model, which could have an impact on our prediction results, our main concern is whether there are any additional correlations among the four predictor variables. The data in the table demonstrate that there is no situation that the absolute value of correlation between the predictor variables is excessively high, which are less than 0.8, indicating that there is no multicollinearity between our predictor variables. Poverty and education had a negative correlation (r = -0.3198). This correlation is somewhat in line with the previously mentioned geographic distribution of bachelors and poverty.

	LNBELPOV100	PCTBACHMOR	PCTSINGLES	PCTVACANT
LNBELPOV100	1.0000	-0.3198	-0.2905	0.2495
PCTBACHMOR	-0.3198	1.0000	0.1975	-0.2984
PCTSINGLES	-0.2905	0.1975	1.0000	-0.1514
PCTVACANT	0.2495	-0.2984	-0.1514	1.0000

Pearson's Correlation of Predictors

Table 2

Table 2. The correlation between each predictor variable and itself.

3.2 Regression Results

The final regression equation 2 we used is represented below:

$$y = \beta_0 + \beta_1 LNBELPOV100 + \beta_2 PCTBACHMOR + \beta_3 PCTSINGLES + \beta_4 PCTVACANT + \varepsilon \quad (2)$$

We regressed the y, which is natural log of Median House Value (LNMEDHVAL) on the natural log of number of households with income below 100% poverty level (LNBELPOV100), percentage of residents in Block Group with at least a bachelor's degree (PCBACHMOR), percentage of housing units that are detached single family (PCTSINGLES), and the percentage of vacant housing units (PCTVACANT). The output of our regression is shown in Table 3 and 4 below. It indicates that these predictors are positively associated with the Median House Value, as the derived p-value for all the 4 variables are lower than 0.0001. The coefficients will have a modified interpretation because we have used a logarithmic transformation for the dependent variable. According to the derived coefficient estimates in Table 3, and noted that when both dependent variable and predictor are transformed, and β_i is small (< 20 in absolute value), we can use the following approximation: $\left(1.01^{\beta_i} - 1\right) \cdot 100\% \approx \beta_i\%$. Since $|\beta_1| = 0.0789 < 20$, we can say that a 1% increase in the number of households with income below poverty line corresponds to an approximately $\beta_1\% = -0.0789\%$ change in the median house value. Also, when only the dependent variable is transformed, and β_i is small (≤ 0.3 in absolute value), it happens to be the case that $(e^{\beta_i} - 1) * 100\% \approx 100\beta_i\%$, in a similar way, Since $|\beta_2| < 0.3$, we could use the approximation formula, and say that 1% increase in the percentage of residents in block group with at least a bachelor's degree, the median house value goes up by approximately 2.09% for a one unit. Furthermore, the 1% increase in the percentage of housing units that are detached single family is associated with the 0.30% decrease for a one unit of median house value, and 1% of the percentage of vacant housing units is associated with an increase of 1.92% increase for a one unit of median house value.

As well as that, the p-value for LNBELPOV100 is less than 0.0001, which indicates that if there is actually no relationship between LNBELPOV100 and the dependent variable LNMEDHVAL (i.e., if the null hypothesis that $\beta_1=0$ is actually true), the probability of obtaining a β_1 coefficient estimate of -0.789 is less than 0.0001. In the

same way, the p-value of PCBACHMOR is less than 0.0001, which represents that if there is actually no relationship between PCBACHMOR and the dependent variable MEDHVAL (i.e., if the null hypothesis that $\beta_2 = 0$ is actually true), the probability of gaining a $\beta_2$coefficient estimate of 0.0209 is less than 0.0001. Also, the same interpretations can be made for the p values of PCTVACANT and PCTSINGLES - both of these predictors are statistically significant with a very low p value < 0.0001. These low probabilities indicate that we can safely reject

$$H_0: \beta_1 = 0 \; for \; H_a: \beta_1 \neq 0;$$

$$H_0: \beta_2 = 0 \; for \; H_a: \beta_2 \neq 0;$$

$$H_0: \beta 3 = 0 \; for \; H_a: \beta 3 \neq 0;$$

$$H_0: \beta 4 = 0 \; for \; H_a: \beta 4 \neq 0.$$

There are more than a half of variance in the dependent variable explained by the model, R-Squared is 0.6623 and Adjusted R-Squared is 0.6615. The low p-value associated with the F-ratio represents that it is feasible for us to reject the null hypothesis that all the coefficients in the model are 0.

term	estimate	std.error	statistic	p.value
(Intercept)	11.1137661	0.0465330	238.836351	0.00e+00
LNNBELPOV100	-0.0789054	0.0084569	-9.330279	0.00e+00
PCTBACHMOR	0.0209098	0.0005432	38.493943	0.00e+00
PCTVACANT	-0.0191569	0.0009779	-19.590280	0.00e+00
PCTSINGLES	0.0029769	0.0007032	4.233544	2.42e-05

Regression Summary

Table 3

```
Multiple R-squared:  0.6623,    Adjusted R-squared:  0.6615
```

Table 3. The regression summary of each variable (PCTVACANT, LNBELPOV100, PCTSINGLES, PCTBACHMOR), showing the estimate coefficient, standard error, t-statistics, and p-value.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
LNNBELPOV100	1	122.659356	122.6593562	913.2481	0.00e+00
PCTBACHMOR	1	273.604325	273.6043251	2037.0939	0.00e+00
PCTVACANT	1	53.074612	53.0746124	395.1618	0.00e+00
PCTSINGLES	1	2.407244	2.4072438	17.9229	2.42e-05
Residuals	1715	230.343542	0.1343111	NA	NA

Regression Summary

Table 4

Table 4. The regression summary of the variables (PCTVACANT, LNBELPOV100, PCTSINGLES, PCTBACHMOR).

3.3 Regression Assumption Checks

In this section, we will focus on testing model assumptions and aptness. According to the histograms of variable distribution, the validity of the assumption that there is a linear relationship between the dependent variable LNMEDHVAL and each of its predictors will be examined. As we can see in the scatter plots below, PCTBACHMORE and LNMEDHVAL appear to have the most significant linear relationship. However the other scatter plots seem not to be exactly linear. Although it doesn’t quiet meet our first assumption, there is no strong polynomial relationship in our case, our liner regression model is appropriate through the observations.

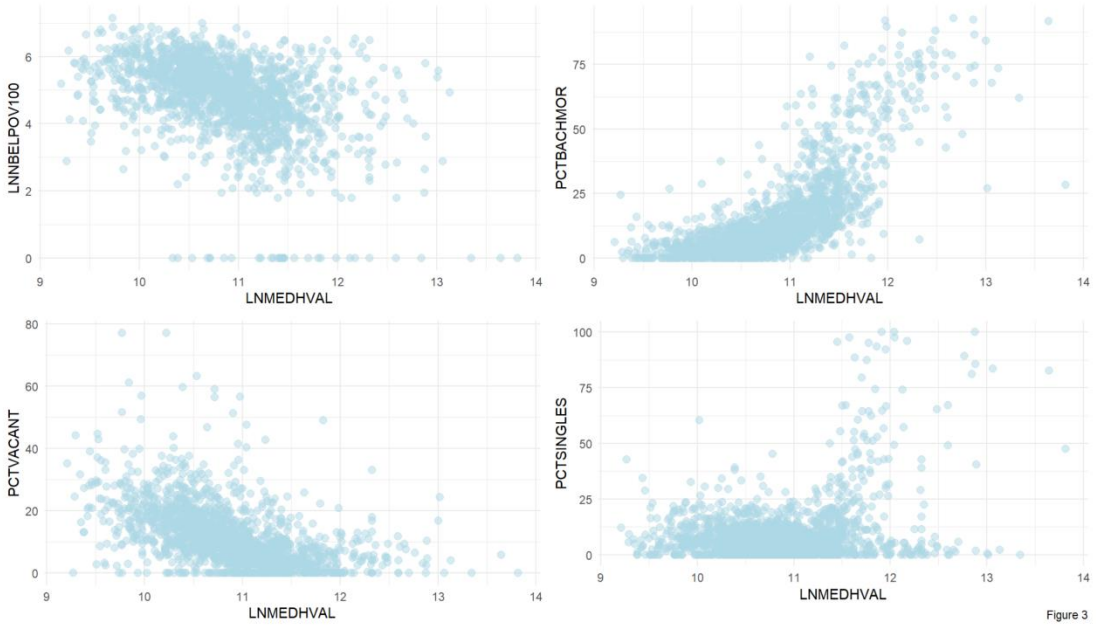


Figure 3. The scatter plot of the relationship between each predictor variable and the natural log of the median house value.

The second assumption of OLS regression is the normality of residuals. The histogram of the standardized residuals is shown in Figure 4 below. After the variable predicts the result, the standardized residual can be understood as the residual divided by the standard error. It is possible to reflect the lateral normality of variable residuals using standardized residuals. The distribution of standardized residuals is normal.

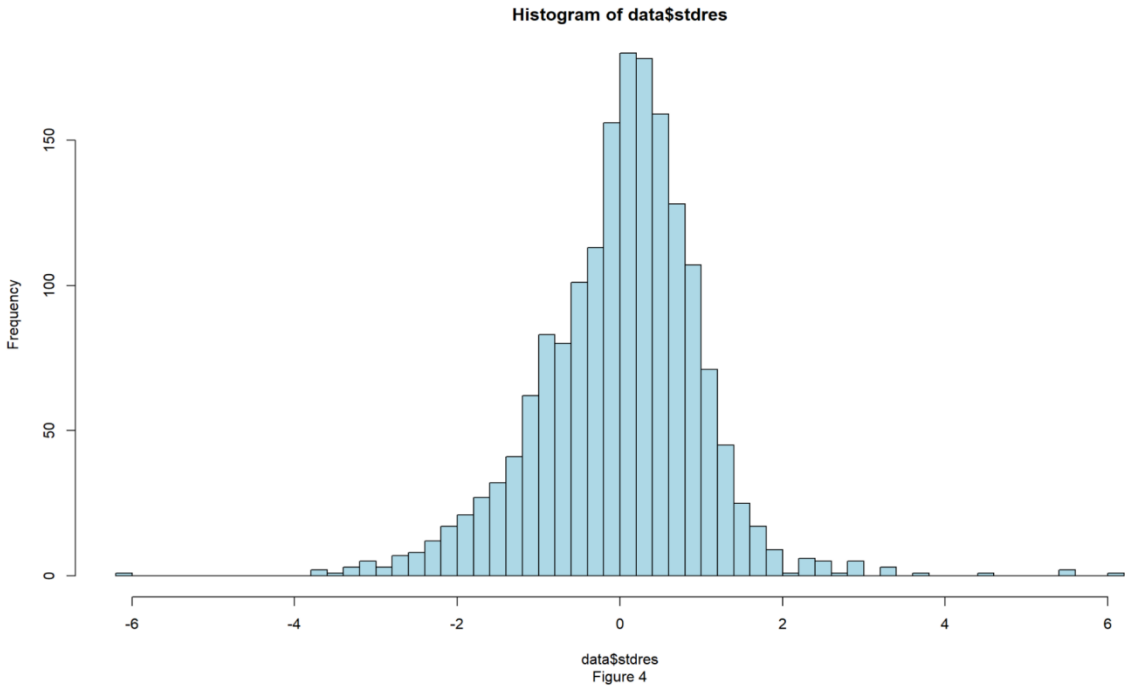


Figure 4. The histogram of the normalized residuals.

The homoscedasticity of the residuals is the third assumption will be investigated in the regression model. Homoscedasticity refers to the random distribution of residual variance regardless of the values of each variable and also indicates the accuracy of model. The scatter plot in Figure 5 depicts the relationship between the standardized residuals and the median house values predicted by the model. Figure 5 demonstrates that the standardized residual distribution matches the dependent variable distribution randomly. What's more, testing the homoscedasticity can help us identify outliers observed in the dataset.

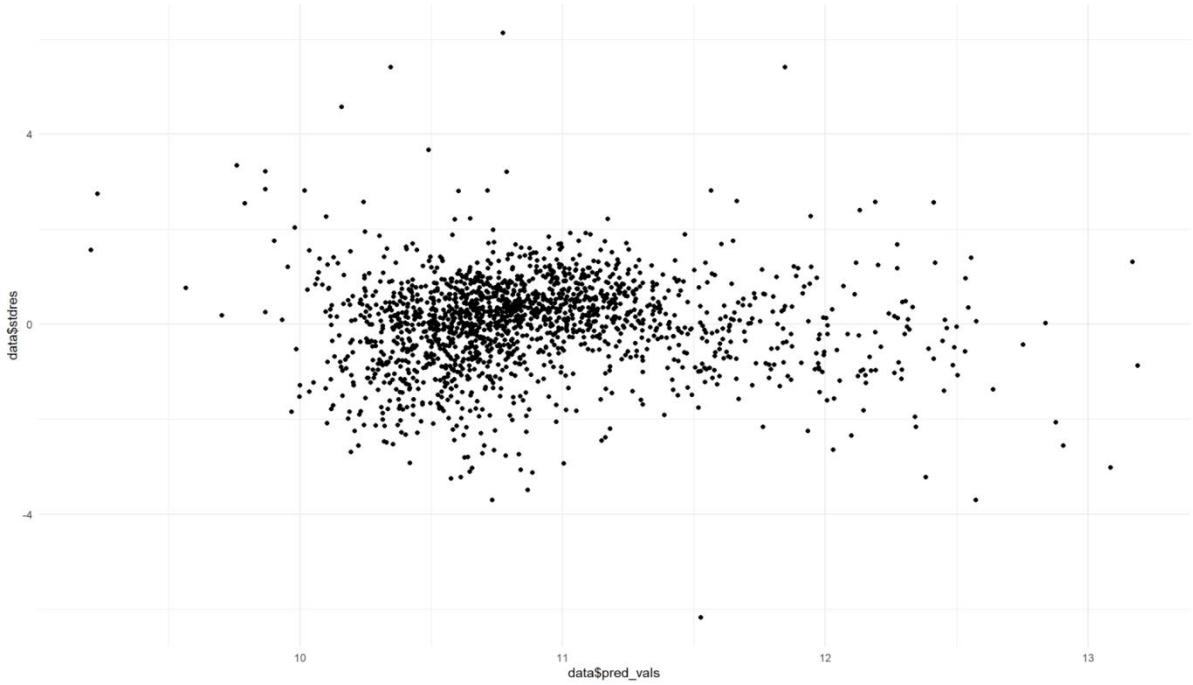
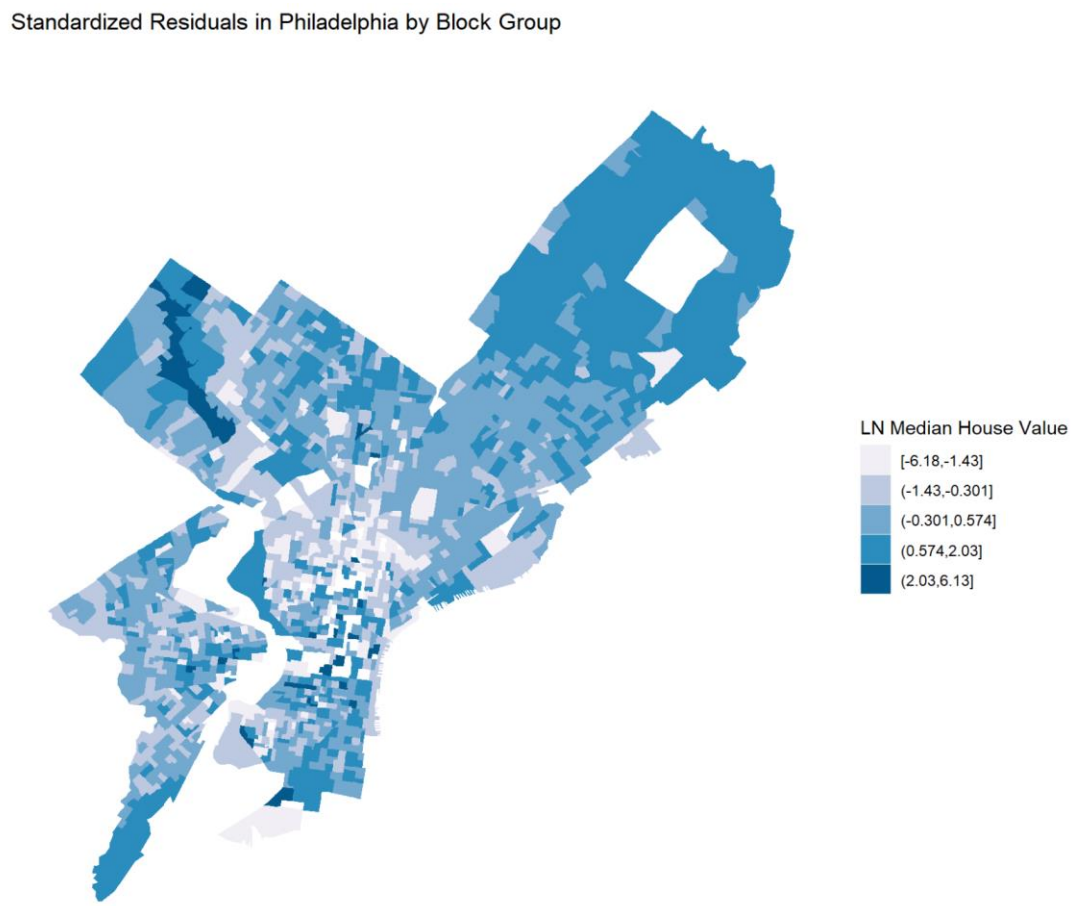


Figure 5. The relationship between the standardized residuals and the median house values predicted in our model.

From the standardized residuals on the map 3 equivalent in geographical space distribution in Philadelphia, we can see the distribution of the data has certain characteristics, it can be seen in the middle of the Philadelphia area standardized residual data aggregation, which shows that our model data has the existence of spatial autocorrelation, according to The First Law of Geography (Waldo Tobler). For an information point in a certain location, its surrounding data information has an impact on it, and the near things are more closely related than the distant things. In the data exploration section, we have shown five maps. We can preliminarily see the spatial related information of the data. Comparing this standardized residual map with the previous maps shows that the spatial distribution of the residuals is more random than the distribution of the predicted results. We will examine the spatial autocorrelation of the variables and residuals and run spatial regressions in the future.



Map 3. A map of standardized Residuals in Philadelphia by block group

3.4 Additional Models

The results of the stepwise regression are shown in Table 5. Due to the lowest AIC, all the four predictors are kept in the final model.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1715	230.3435	-3448.073

Stepwise ANOVA

Table 5

Table 5. The result of stepwise regression.

Tables 6, 7, and 8 show the model summary statistics where PCTVACANT and MEDHHINC are inserted. We have calculated the MSE and RMSE of both regressions, which are displayed in Table 9. These results show that the original model is superior

to the second model because the two indices are relatively lower.

term	estimate	std.error	statistic	p.value
(Intercept)	10.4302971	0.0330137	315.93870	0
MEDHHINC	0.0000210	0.0000007	29.04807	0
PCTVACANT	-0.0186386	0.0012232	-15.23757	0

Regression Summary

Table 6

Table 6. Additional regression summary.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MEDHHINC	1	300.21459	300.2145858	1532.3709	0
PCTVACANT	1	45.48828	45.4882827	232.1837	0
Residuals	1717	336.38621	0.1959151	NA	NA

ANOVA

Table 7

Table 7. Regression summary statistics.

```
Call:
lm(formula = LNMEDHVAL ~ MEDHHINC + PCTVACANT, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.51313 -0.24807 -0.04362  0.20179  2.82016

Coefficients:
                Estimate    Std. Error t value            Pr(>|t|)
(Intercept) 10.4302970977  0.0330136734  315.94 <0.0000000000000002 ***
MEDHHINC     0.0000209907  0.0000007226   29.05 <0.0000000000000002 ***
PCTVACANT   -0.0186385838  0.0012231989  -15.24 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4426 on 1717 degrees of freedom
Multiple R-squared:  0.5068,    Adjusted R-squared:  0.5063
F-statistic: 882.3 on 2 and 1717 DF,  p-value: < 0.00000000000000022
```

Table 8. R-Squared and Adjusted R-Squared.

	MSE	RMSE
Regression 1 (4 predictors)	0.134	0.366
Regression 2 (2 predictors)	0.196	0.443

Table 9. Comparison of regression 1 and 2 with difference in MSE and RMSE values.

4 Limitations & Conclusion

The goal of this regression project is to predict median house value in Philadelphia with four predictor variables by examining their relationships with median house price. As the research illustrated, the median house value is examined along with the local factors that may affect it including number of households with income below 100% poverty level, percentage of residents in Block Group with at least a bachelor's degree, percentage of housing units that are detached single family, and the percentage of vacant housing units. The model of independent variables for predicting the result of contribution and influence examines the distributional characteristics of each variable, calculates correlations between independent variables. As can be seen from the maps, house value is highly correlated with educational background of the local residence. The regression model shows that education level has a greater impact to the prediction result. Additionally, a better prediction can be obtained by using the natural logarithm of predictor variables. This model's ability to predict the median house price can be confirmed by the prediction process and result test.

According to the result of the prediction, the p-values of all four predictor variables are very small, which shows that they have an impact on the prediction of the dependent variable. The R-Squared is 0.6623. Since it is possible that the absence of a predictor variable has no impact on the outcome of the prediction, further investigation of the prediction model is required to ascertain the significance of the predictor variables in the model. To test whether the removal and reduction of variables will have a clear effect on the forecasting model, we build an additional model using just two predictor variables: the percentage of vacant units and the median household income. Subsequently, we added cross-validation with 5 folds, and we used root mean square error as one of the key criteria for comparing the prediction outcomes between the two models. The 4 predictors model is better than 2 predictors model.

We did not use the number of people living below the poverty line as it was felt that the raw results did not accurately reflect how many people in the block group were living. This is a natural variable. When the variance of the dependent variable (housing values) increases with the independent variable, log transformation can be effective in addressing this issue; otherwise, the predicted results would be more accurate.

Finally, the data in the space's spatial geographic neighborhood autocorrelation is not taken into account by our model. More variables are required for the model to explain the relationship between the house values in the space, as suggested by the standardized residuals distribution map.

5 References

HOEKSTRA, J. and VAKILI-ZAD, C. (2011). HIGH VACANCY RATES AND RISING HOUSE PRICES: THE SPANISH PARADOX. *Tijdschrift voor economische en sociale geografie*, 102(1), pp.55–71. doi:10.1111/j.1467-9663.2009.00582.x.

Kaushal, A. and Shankar, A. (2021). House Price Prediction Using Multiple Linear Regression. *SSRN Electronic Journal*. doi:10.2139/ssrn.3833734.

Painter, G. and Yu, Z. (2008). Leaving Gateway Metropolitan Areas in the United States: Immigrants and the Housing Market. *Urban Studies*, 45(5-6), pp.1163–1191. doi:10.1177/0042098008089864.

WHITTINGHAM, M.J., STEPHENS, P.A., BRADBURY, R.B. and FRECKLETON, R.P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, 75(5), pp.1182–1189. doi:10.1111/j.1365-2656.2006.01141.x.