Regression Analysis of Houses in Saratoga County (2006)

Mary Keonoupheth & Sara Rettus

Abstract

The objective of this project was to apply methods and analysis used in linear regression with the purpose of fitting the best model for the data set Houses in Saratoga County (2006), in which home price was the response variable. The data set was collected in 2007 for a case study titled *"How much is a Fireplace."* The source of the data, minus the variable for college graduates, was extracted from the Saratoga County New York assessment database. The context of the data is important to understanding the limitations of the model. While we concluded there is a significant relationship between many of the regressors and home price, with an adjusted R-squared score of approximately 0.65, it is not a practical model for predicting assessment home prices.

There were 1728 observations in our data set. The set contained sixteen variables: price (response variable), lot size, age, land value, living area square feet, percentage of neighborhood that graduated college, number of bedrooms, number of fireplaces, number of bathrooms, total number of rooms in the home, e.g., bedrooms, bathrooms, kitchen, living room, etc., type of heating (electric, hot air, and hot water/steam), fuel type (electric, gas, and oil), sewer (septic, none, public/commercial), waterfront, new construction, and central air. Six of these variables were categorical, of which three had two levels.

Prior to selecting variables, the full model was created, transformed and analyzed. While the variance for the full model, without transformation, was stable, the distribution of the residuals was not normal. After performing a number of response transformations, the square root transformation did the best job of lessening the positive skew of the residuals.

The model was further analyzed for influence points and multicollinearity. Severe variance inflation was found for electric home fuel and electric heating. Our variable selection consisted in using backward selection. The method produced the model in which college graduates, fireplaces, sewer and fuel type were deleted. Deleting variables did not change the skewed distribution of the residuals. After performing a number of response transformations, the square root transformation of the response yielded the best results. The response transformation changed the significance of bedrooms to our model; as such, we re-ran the transformed data set through the variable selection process. Each method provided us with the same results as last time with the exception of dropping the variable, bedroom.

The final analysis on outliers, leverage and influence showed that the number of influential observations either remained the same or increased.

Last we checked for multicollinearity. The main source of multicollinearity in the previous model had been fuel type and heating. With fuel type no longer a part of our model, all of our

variance inflation factors were lowered. Our condition number lowered from 255 to 16 such that multicollinearity was no longer a factor in our model.

Introduction

The objective of this project is applying methods and analysis used in linear regression with the purpose of fitting the best model for the data set, Houses in Saratoga County (2006), in which home price was our response variable.

Data for this project was collected by Candice Corvetti, in 2007, for the case study *"How much is a Fireplace?"* (Pruim, 2020). The source for this dataset, except for the variable percentage of college graduates in the area, was collected from the Saratoga County New York assessment database (Real Property Tax Services).

Data source plays an important role in understanding this dataset and the findings of our model. Home prices were not collected from home sales, but instead taken from the home value listed in the assessment database. This value is meant to reflect fair market value, as determined by a county assessor for the purpose of calculating yearly property taxes but does not necessarily reflect actual market values (The 2019 Tax Resource). Moreover, absent from the dataset, but part of each home assessment, were variables such as school district, home improvements, overall condition of the home, and home grade. Given the absence of this information, it is not surprising that model regressors were only able to account for around 65% of total model variance.

As such, we conclude that while the model is not practical for predicting future assessment of home values or actual market value, it does show that home value has a relationship to lot size, land value, age of home, square footage of the living area, number of bathrooms, total number of rooms, hot air heating, waterfront property, new construction classification, and central air.

Data Set Observations and Variables

The number of observations in our data set numbered 1728. The data set contained sixteen variables including the response variable, price. The other fifteen regressors were lot size, age, land value, living area square feet, percentage of neighborhood that graduated college, number of bedrooms, number of fireplaces, number of bathrooms, total number of rooms in the home, e.g., bedrooms, bathrooms, kitchen, living room, etc., type of heating (electric, hot air, and hot water/steam), fuel type (electric, gas, and oil), sewer (septic, none, public/commercial), waterfront, new construction, and central air. Six of these variables were categorical, of which three had two levels.

Data Exploration

Data exploration included viewing the summary of the dataset and searching for inconsistencies.

Reviewing the quantiles along with the minimum and maximum values for our numeric variables provided a big picture view of our data ranges. Summaries also provided clues to possible data defects, such as the minimum values of 0 for bathrooms and lot size. We were also able to detect possible pricing errors, such as the minimum value for home price listed as $5000, when the median price was $189000 with a maximum of $775000.

Queries were run on homes with land value greater than the home price, homes with less than one bathroom, homes with no sewer system, and homes in which the lot size was zero.

The search yielded three homes whose land values were greater than home price and twelve homes were listed with no sewer system. The query showed that only one home was missing a bathroom and two homes had a lot size of zero. Since these observations accounted for less than 1% of the original dataset, we decided to leave these observations in the set as they will have little effect on the model.

Model Building

To provide a basis with which to understand any model reduction a full analysis was conducted on the full model.

Full Model Equation

$$\hat{y} = 9259 + 7599\text{lotSize} - 130.4\text{age} + 0.9219\text{landValue} + 69.96\text{livingArea} - 110.2\text{pctCollege} - 7835\text{bedrooms} + 1037\text{fireplaces} + 23110\text{bathrooms} + 3020\text{rooms} - 82.45\text{heatinghotair} - 10370\text{heatinghotwater/steam} + 10930\text{fuelgas} - 6550\text{fueloil} + 3321\text{sewerpublic/commercial} + 4845\text{sewerseptic} + 120200\text{waterfrontYes} - 45440\text{newConstructionYes} + 9953\text{centralAirYes}$$

Initial Findings

H_0: The regressors do not have an influence on the response home price.

H_1: At least one the regressors have a significant relationship to the response home price.

Significance level: alpha = 0.01

F-statistic was 179 on 18 and 1709 degrees of freedom with a p-value: < 2.2e-16. Adjusted R-squared value was 0.6498, with a residual standard error of 58260 on 1709 degrees of freedom.

Conclusion is to accept the alternative hypothesis that at least one regressor has a significant relationship to the response variable home price. Regressors in this model account for 64.98% of the total variance.

P-values for six of the regressors was above the 0.01 significance level, specifically, age, percent of college graduates, fireplaces, heating, fuel, and sewer.

Full Model Adequacy Check

Variance appears stable judging from externally studentized residuals plotted against fitted values shown in Figure 1. However, the normal QQ plot in Figure 2 shows the residuals are not normal. There are several outliers on the upper and lower ends, and the residuals are positively skewed.

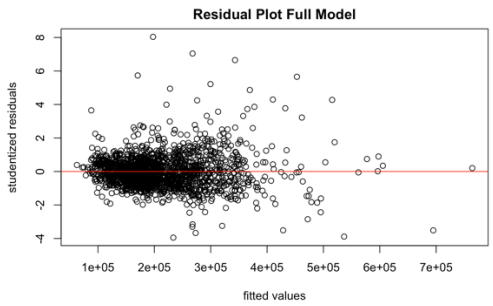


Figure 1 Externally studentized residuals plotted against fitted values

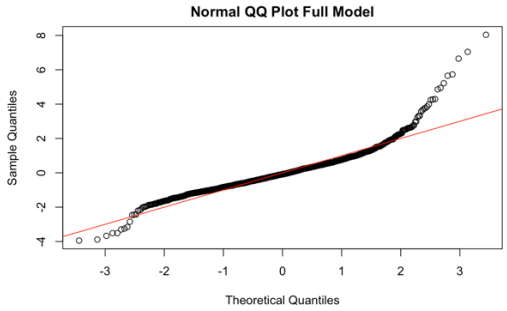


Figure 2 Normal QQ Plot

There are some outliers with values above the absolute value of 3. The residual vs. leverage plot in Figure 3 shows that there are some influential points with high leverage and large residuals in the data set. The cook's distance plot shows that there are no influential points shown in Figure 4 since no values exceed 1.

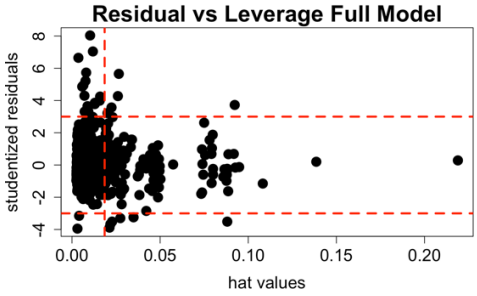


Figure 3 Residual vs. Leverage Plot

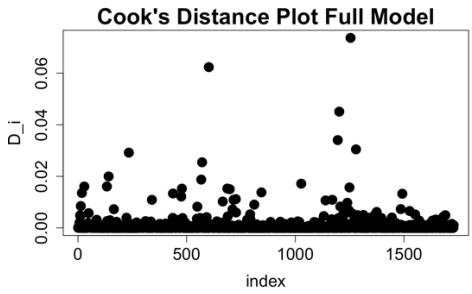


Figure 4 Cook's Distance Plot

The DFFITS plot in Figure 5 shows that there are many points that have influence on their fitted values. The covratio plot in Figure 6 shows that there are many points that have positive influence on the precision of coefficient estimates, while there are many points that have negative influence on the precision of coefficient estimates.

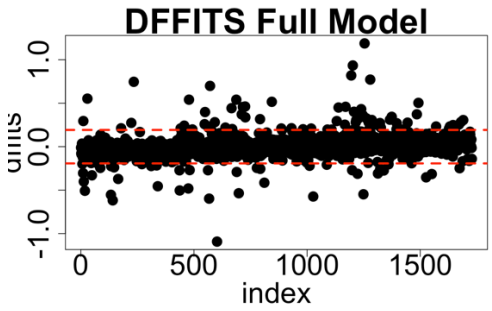


Figure 5 DFFITS Plot

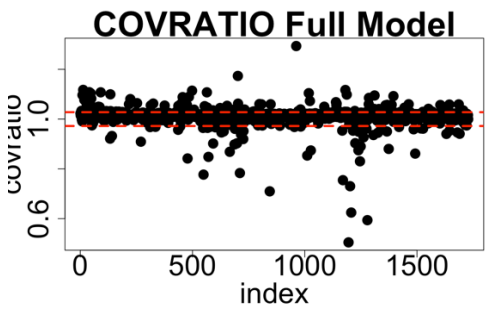


Figure 6 Covratio Plot

The DFBETA plots show that observation 1254 shown in Figure 7, 8, and 9 has some influence on the intercept and regressors 14 (new construction) and 15 (central air). The DFBETA plot shown in Figure 10 shows that observation 1202 has influence on regressor 7 (fireplaces). These observations should be monitored since they also appear as influential points in the leverage vs. residual plot.

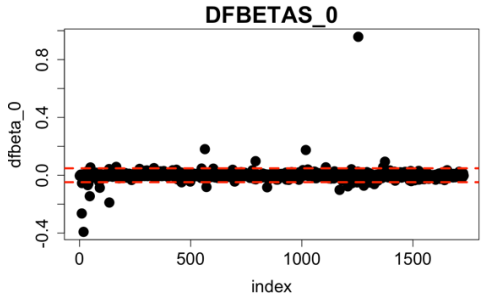


Figure 7 DFBETA plot on intercept

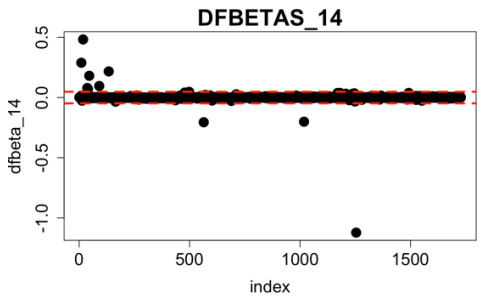


Figure 8 DFBETA plot on regressor 14 (new construction)

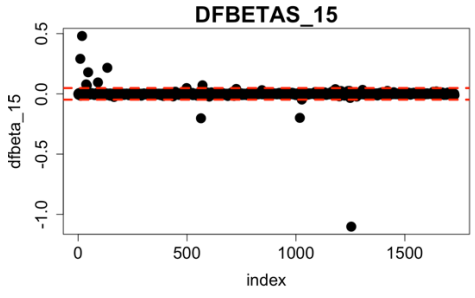


Figure 9 DFBETA plot on regressor 15 (central air)

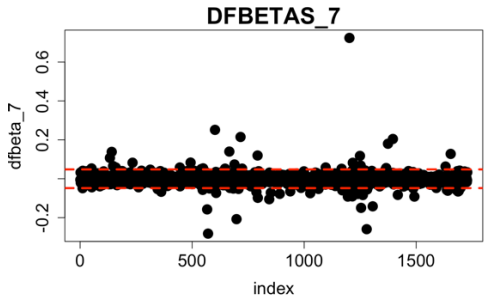


Figure 10 DFBETA plot on regressor 7 (fireplaces)

Transformations

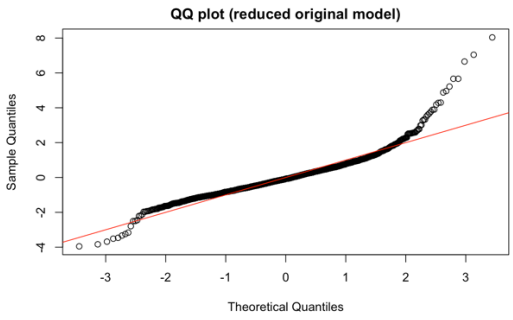
Three different transformations were attempted on y to counteract the skewed nature of the residuals: square root of the response, box cox lambda = 1/4, and log transformation. The log transformation of y increased the normality of the top of the plot but decreased it for the bottom. Moreover, the residuals plotted against fitted value had a bit more curvature than the square root transformation. The box cox and square root transformations had adjusted R-squared values, 0.6231 and 0.6453 respectively, with both lessening, but not fully correcting, the skewed nature of the residuals. The square root transformation of y was selected due to its slightly higher adjusted R-squared value and its simplicity.

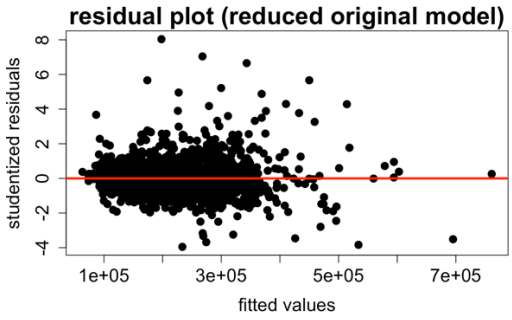
Multicollinearity

Moderate correlation was found between regressors, rooms, living area, bathrooms, and bedrooms. This correlation is expected since these regressors are all some form of room. There were high VIF scores for each type of heating, fuel, and sewer indicating multicollinearity among these variables. The condition number was 254.7621, which indicates moderate multicollinearity.

Building the Reduced Model

Backward elimination was used for model selection on the full model, square root model, and box cox model. This method helped eliminate the insignificant regressors: percent college, bedrooms, fireplaces, fuel, and sewer, for a better and simpler model. After backward elimination, the original model had the greatest adjusted R-squared score of 0.6506, while the square root and box cox model both had an adjusted R-squared of 0.646. However, when comparing their residual plots and QQ norm plots, we determined that the reduced square root model was the best model overall with its residuals slightly more normally distributed than the other models shown in Figure 11.





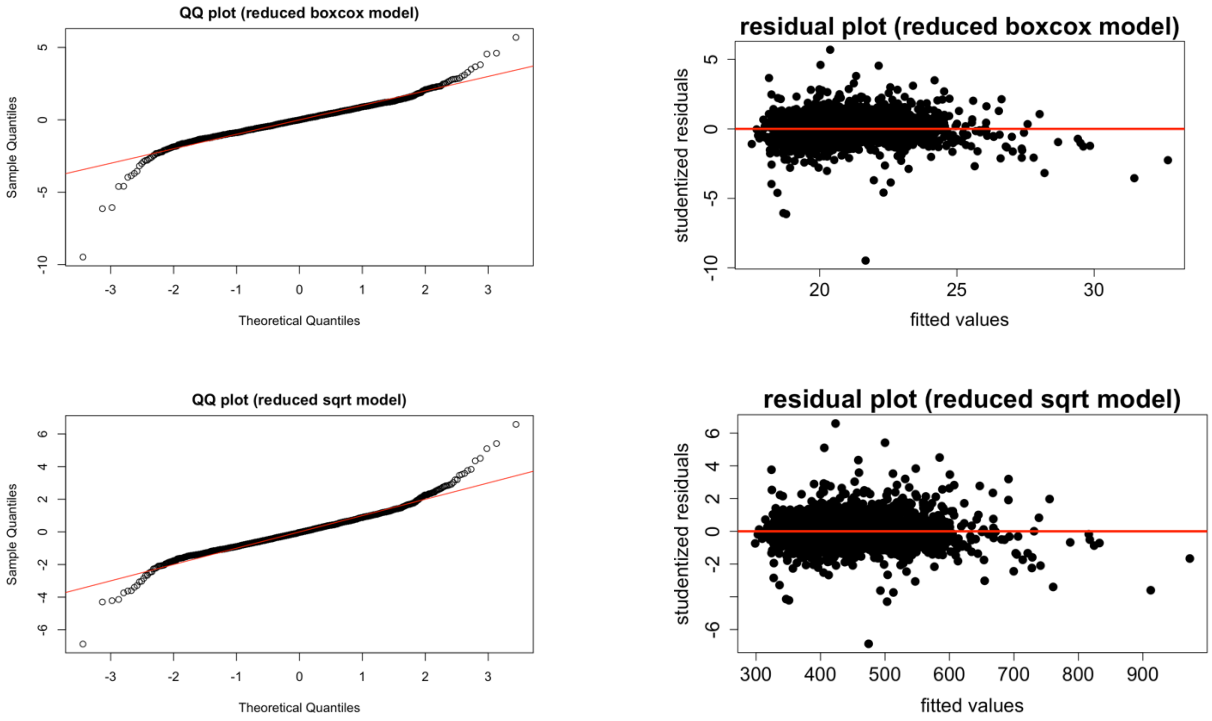


Figure 11 Residual plots and QQ norm plots for original model, square root transformation, and box cox transformation of the response variable after backward elimination was performed.

Final Equation

Reduced model with square root transformation of y:

$\sqrt{\hat{y}}$ = 229.5 + 8.280lotSize - 0.2287age + 0.0008674landValue + 0.06698livingArea + 24.72bathrooms + 2.171rooms + 13.08heatinghotair + 4.516heatinghotwater/steam + 115.8waterfrontYes - 41.30newConstructionYes + 10.77centralAirYes

The F-statistic for the final model is 287.5 on 11 and 1716 DF, p-value: < 2.2e-16 with an adjusted R-squared score of 0.646.

Conclusion is that these regressors, either individually or in conjunction with other variables, have a significant relationship to the response variable price. The final model has significantly smaller coefficients than the initial model, with waterfront having the largest non-intercept coefficient.

Final Model Adequacy Check

Variance still appears stable judging from externally studentized residuals plotted against fitted values shown in Figure 12. The QQ norm plot in Figure 13 shows that the normality issue has improved but is still not perfectly normal. There are still several outliers on the upper and lower ends, and the residuals are positively skewed.

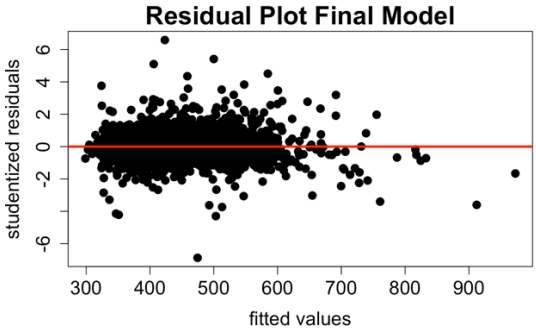


Figure 12 Externally studentized residual plot against fitted values

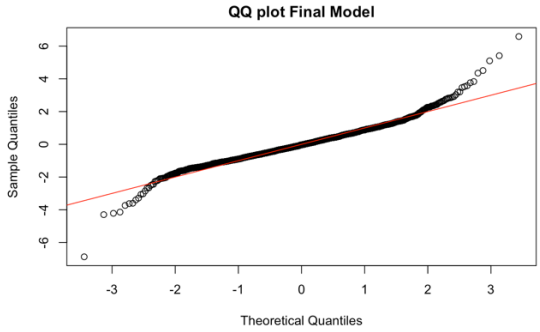


Figure 13 QQ norm plot

There are some outliers with values above the absolute value of 3. The residual vs. leverage plot for the final model in Figure 14 shows that there are very few influential points with high leverage and large residuals in the data set. The cook's distance plot in Figure 15 shows that there are no influential points in the final model.

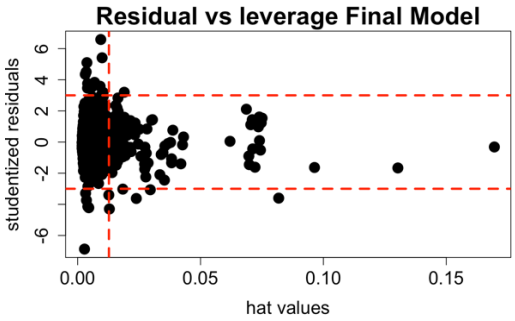


Figure 14 Residual vs. Leverage plot

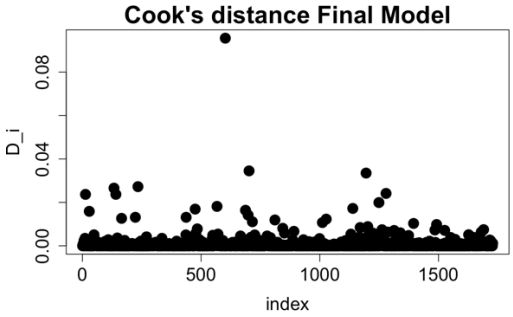


Figure 15 Cook's distance plot

The DFFITS plot in Figure 16 shows that there are many points that have influence on their fitted values. The covratio plot in Figure 17 shows that there are many points that have positive influence on the precision of coefficient estimates, while there are many points that have negative influence on the precision of coefficient estimates.

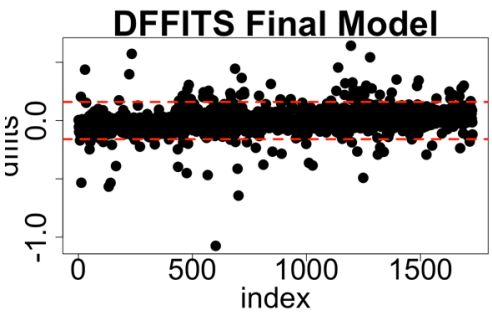


Figure 16 DFFITS plot

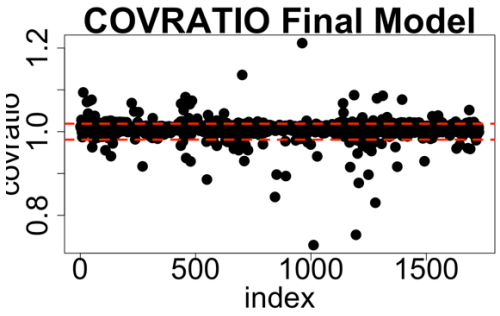


Figure 17 Covratio plot

Multicollinearity for Final Model

With fuel no longer a part of the model, the worst of the multicollinearity from the full model was removed. Correlation for the new reduced model still showed moderate correlation between living area/bathrooms and living area/total rooms. Since the living area increases as rooms increase, some correlation is expected. All the VIF scores for each of the variables decreased from the full model. These values indicate no issue with multicollinearity and do not warrant additional transformation. Last, the new condition score, 16.45352, is significantly lower than the initial value of 254.7621 for the full model. As such, we conclude that the reduced square root model does not have an issue with multicollinearity.

Concluding Summary:

In analyzing the dataset Houses in Saratoga County (2006) we concluded that lot size, age, land value, living area, bathrooms, hot air heating, waterfront, new construction, and central air had a significant relationship to the response variable home price. The final equation for our model

$\sqrt{\hat{y}}$ = 229.5 + 8.280lotSize - 0.2287age + 0.0008674landValue + 0.06698livingArea + 24.72bathrooms + 2.171rooms + 13.08heatinghotair + 4.516heatinghotwater/steam + 115.8waterfrontYes - 41.30newConstructionYes + 10.77centralAirYes

was built after performing a square root transformation on our response variable and running our dataset through backward selection to eliminate insignificant variables. While the variance for this model is stable, it does suffer some adequacy issues in that the distribution of the residuals maintains a somewhat positive skew, and the data contains several influential observations. Multiple attempts at response transformations only minimally lessened the skewed nature of the residuals. While we conclude that there is a significant relationship between the response variables and home price, the adjusted R-squared value at 0.646 results in an even lower prediction R-squared value, which makes our model impractical for the purposes of predicting home assessment values.

Works Cited:

National Register of Historic Places listings in Saratoga County, New York. (2020, October 19). Retrieved December 07, 2020, from https://en.wikipedia.org/wiki/National_Register_of_Historic_Places_listings_in_Saratoga_County,_New_York

Real Property Tax Services. (2020, October 27). Retrieved December 06, 2020, from https://www.saratogacountyny.gov/departments/real-property-tax-service-agency/

Randall Pruim, D. (2020, September 13). SaratogaHouses: Houses in Saratoga County (2006) in mosaicData: Project MOSAIC Data Sets. Retrieved December 06, 2020, from https://rdrr.io/cran/mosaicData/man/SaratogaHouses.html

The 2019 Tax Resource. (n.d.). Retrieved December 06, 2020, from http://www.tax-rates.org/new_york/saratoga_county_property_tax