NEW YORK CITY HOME PRICE PREDICTION

MGT-6203 Group Project Final Report

APRIL 16, 2023

TEAM 3
Xiao Yi Xu, Inhyuck Lee, Lin Lin, Kihyup Yoon, Omar El-Koussy

Table of Contents

Abstract

This project consists of a detailed analysis of factors influencing NYC housing prices. Housing is the largest asset that an individual obtains, and a lot of time and effort are used to purchase and sell homes. The problem of housing prediction is that it has high complexity as it involves many variables, both internal and external. To increase the accuracy in housing price prediction, Team 3 utilized the NYC housing data from Zillow, Crime Data from NYPD, and population per zip code to build a NYC housing price prediction model. The research focused on the following: What are the best predictors for NYC housing prices, which neighborhoods have the highest variation of property prices, and is housing listing intentionally set to trigger a bidding war? The approach to modeling included data cleaning, data transformations, linear regression, Principal Component Regression, and variable selection. The researchers found that both internal factors such as bedrooms, bathrooms, fireplace, and external factors such as HOA, property tax rate affects housing price in NYC. These findings can be a significant consideration for both sellers and buyers.

Objective/Problem Statement

Information asymmetry is a well-known feature of the real estate market. Property pricing is not solely determined by the internal factors of property itself (Ockey, 2018), but also depends on external factors such as neighborhood crime rate (Ceccat & Wilhelmsson, 2011) and the macroeconomic market conditions, such as interest rate (RBC, 2011). Thus, homeowners set expectations by comparing housing features, neighborhoods, and nearby sales, despite the limited amount of available data. On the other hand, buyers would also like some transparency on housing prices to make sure that the homes they visit are within their budget, and not to overbid on homes. To bring this transparency into the housing market, our group proposes to develop a model that can predict the market value of a property. While homeowners can use the model to set expectations of the selling price, buyers can clearly understand whether a property is expected to be within their budget.

Business Justification/Impact

An average US citizen has 25% to 40% of their net worth tied to real estate. Homeowners should have a fair expectation of their property value before listing it on the market. This expectation can help them avoid panic selling and help ensure that they do not miss quality offers. Home buyers should also have a clear understanding of property market value, so they can focus their search effort on those properties within their budget. If they enter a bidding war, they can also use this value as a benchmark and avoid overbidding.

Since there are too many variables to consider in housing prices such as location, environment, local school quality and so on, it is often difficult to appraise the housing value. Our model will aim to increase real estate efficiency by setting the baseline, considering certain internal and external factors. Both sellers and buyers will be able to understand the market value, set a reasonable price for housing, and hopefully making quicker real estate deals.

Research Questions

Given a property listing in New York City, we want to accurately predict its final sales price. Our primary research question focuses on what the main factors are that impact property prices in NYC. The following are our supporting research questions:

1. What are the best predictors for NYC housing prices? The predictors can be internal (such as square footage, number of bedrooms) or external (such as neighborhood crime rate).
2. Which neighborhoods have the highest variation of property prices?
3. Given a single active listing, does the listing price represent the true property value or is it intentionally set low to trigger a bidding war?

Hypothesis

We anticipate that both internal and external factors will influence housing prices and that our model will reveal significant correlations between variables such as the number of bedrooms and lot size. Based on prior real estate knowledge, we hypothesize that living area, bathrooms, bedrooms, number of felonies, having a private pool, and school ratings are some of the most important predictors of housing price. For school ratings, both families with and without children should see positive correlation between ratings and housing prices. We expect our selected variables to be powerful predictors of housing prices but acknowledge that not all buyers bid rationally. We anticipate an adjusted R-squared of 0.5 to 0.8. Finally, we expect some new listings to be intentionally underpriced to stimulate bidding wars.

Methodology/Approach

The initial step in the data analysis process involves filtering the data to limit the scope of the analysis. Specifically, the model will be built based on single-family homes that have been sold or recently sold in NYC. Homes that are listed for sale will be put aside as a separate group to investigate later. The treatment of missing data will vary depending on the column type. For continuous variables, such as the number of bedrooms, the team will attempt to impute an average value. If this is not feasible, the data will be marked as missing. For binary variables, such as whether the home has a fireplace, the missing values will be treated as default, false in this instance. The data clean up and analysis will be done via a combination of Python and R.

The main model is a linear regression model with property sales price as the dependent variable and property features as the independent variables. Outliers will be removed, and data normalization may be done. All variables will be used in the cleaned data set, but with consideration for correlated variables. Modeling options include a decision tree model, greedy variable selection, and LASSO regression.

To ensure linear regression assumptions are met, data transformations may be needed. Scatter plots can check for linearity, while residuals can be used to check for constant variance. Interaction terms, log or box-cox transformations can help meet these assumptions. Multicollinearity is also a concern, and variance inflation factors can be used to check and remove highly correlated predictors.

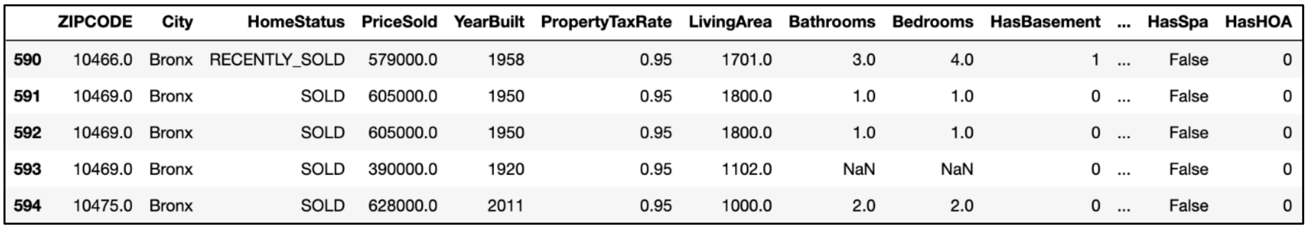
To compare models the team proposes to split all sold homes into training, validation, and testing data with a 60:20:20 ratio. The best model will be selected based on their performance on validation data. The selected model will then be evaluated based on performance on test data. This model will also be used to predict the sales price of new listings, which will be used to answer supporting research question 3.

Data Overview

Three different data sets (using zip code as the key) will be combined to gain insight into different predictors that determine the housing prices in NYC. They are a Zillow data set, an American Community Survey population data set, and a NYPD crime data set all based on year 2021.

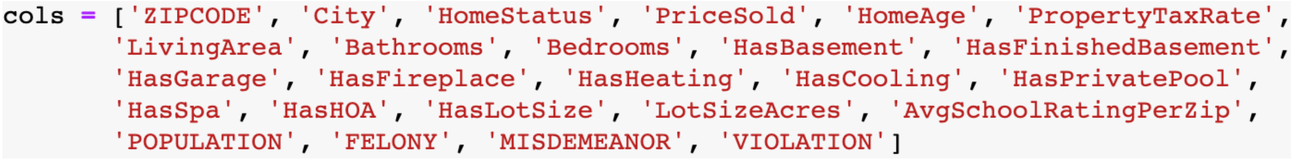
Data Cleaning

The Zillow data set contains 75k+ observations and 1.5k+ predictors, however, we will only be using a subset. Based on the team's judgement on which predictors will add value to our NYC regression model, we manually eliminated and landed on less than 50 predictors to work with. The population data set provides a population count per zip code and did not need any clean up. For the crime data set, we extracted zip codes given the latitude and longitude of the crime using GeoPy API and counted the number of felonies, misdemeanors, and violations per zip code. We then merged these 3 data sets using Python. Furthermore, please see Appendix section "Steps Taken for Data Clean Up" for the steps we took to clean up the data, including removing/creating new predictors, removing outliers, and mean imputations. The cleaned dataset contains 20k+ observations and 25 predictors. 'PriceSold' is the dependent variable and ultimately what we want to predict. All other predictors are independent variables. See Figures 1 and 2 below for sample screenshots of the clean data set along with our predictors list. More detailed screenshots can be found in the Appendix under Data Sets.



	ZIPCODE	City	HomeStatus	PriceSold	YearBuilt	PropertyTaxRate	LivingArea	Bathrooms	Bedrooms	HasBasement	...	HasSpa	HasHOA
590	10466.0	Bronx	RECENTLY_SOLD	579000.0	1958	0.95	1701.0	3.0	4.0	1	...	False	0
591	10469.0	Bronx	SOLD	605000.0	1950	0.95	1800.0	1.0	1.0	0	...	False	0
592	10469.0	Bronx	SOLD	605000.0	1950	0.95	1800.0	1.0	1.0	0	...	False	0
593	10469.0	Bronx	SOLD	390000.0	1920	0.95	1102.0	NaN	NaN	0	...	False	0
594	10475.0	Bronx	SOLD	628000.0	2011	0.95	1000.0	2.0	2.0	0	...	False	0

Figure 1: Sample Screenshot of Clean Data Set



```
cols = ['ZIPCODE', 'City', 'HomeStatus', 'PriceSold', 'HomeAge', 'PropertyTaxRate',
        'LivingArea', 'Bathrooms', 'Bedrooms', 'HasBasement', 'HasFinishedBasement',
        'HasGarage', 'HasFireplace', 'HasHeating', 'HasCooling', 'HasPrivatePool',
        'HasSpa', 'HasHOA', 'HasLotSize', 'LotSizeAcres', 'AvgSchoolRatingPerZip',
        'POPULATION', 'FELONY', 'MISDEMEANOR', 'VIOLATION']
```

Figure 2: Predictor List of Clean Data Set

Data Preparation

Since the cleaned data set had 25 predictors, initial exploratory data analysis would be necessary to identify which predictors could be removed from the model before moving on to modeling and feature engineering. The continuous predictors were plotted on scatter plots and the categorical predictors were plotted on boxplots against price sold and the log of price sold. This process also helped update the data cleaning process by identifying obvious outliers. Based on these initial graphs, the price sold does not appear to be strongly correlated with any predictor. Bathrooms and bedrooms seem to have some sort of positive quadratic relationship with price. A correlation table of the predictors was created to check if interdependence or multicollinearity is present between any two predictors. The table revealed that 'HasBasement' has a strong correlation with 'HasFinishedBasement," with a correlation of 0.78. (Note that in NYC, areas below the main entry, like the basement, is not included in the official square footage and this is already considered in the original Zillow data set.) A house that has cooling is also correlated with a house that has heating with a correlation value of 0.55. The crimes in an area were also moderately correlated with each other and highly correlated with population, indicating that we can consider removing some of them from the analysis. The scatter plots for the continuous predictors and box plots for the categorical variables against Price Sold/log of Price Sold can be seen below in Figures 3-5:

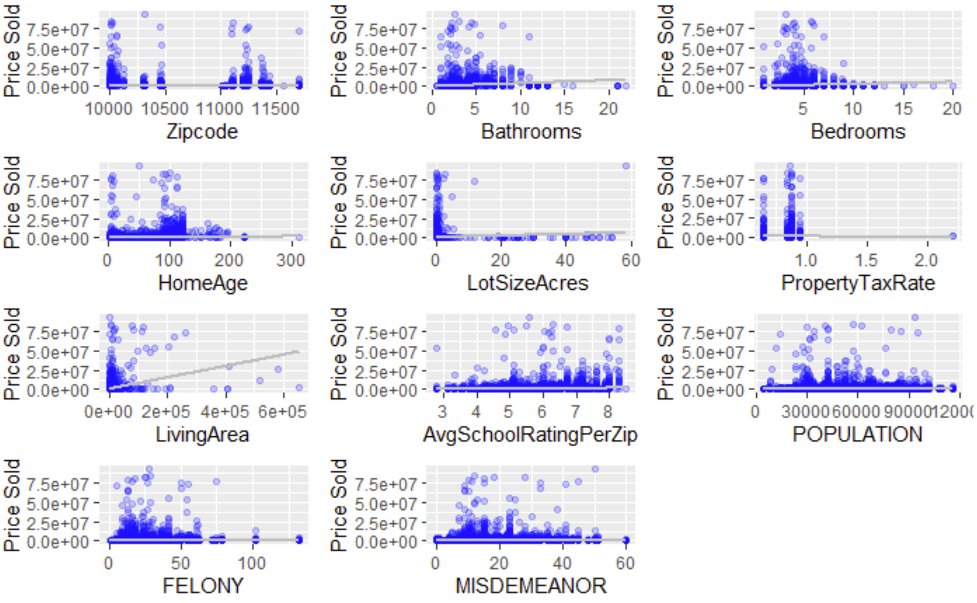


Figure 3: Scatter plots for Price Sold vs. continuous predictors

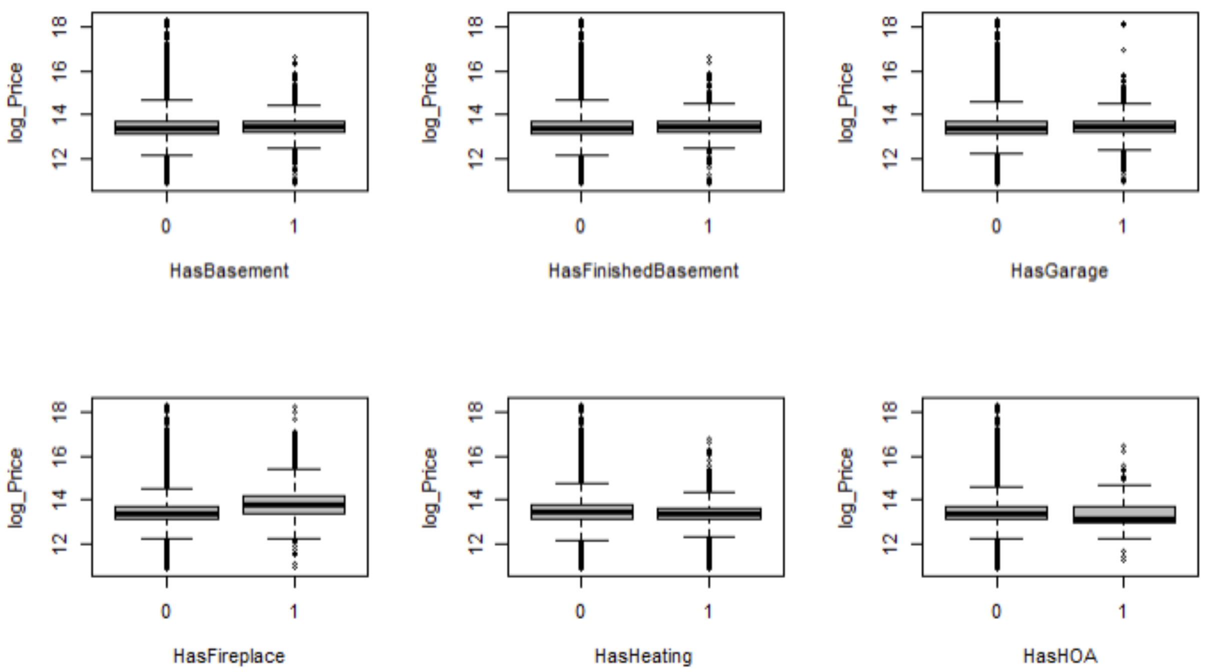


Figure 4: Box plots for log(Price Sold) vs. categorical predictors

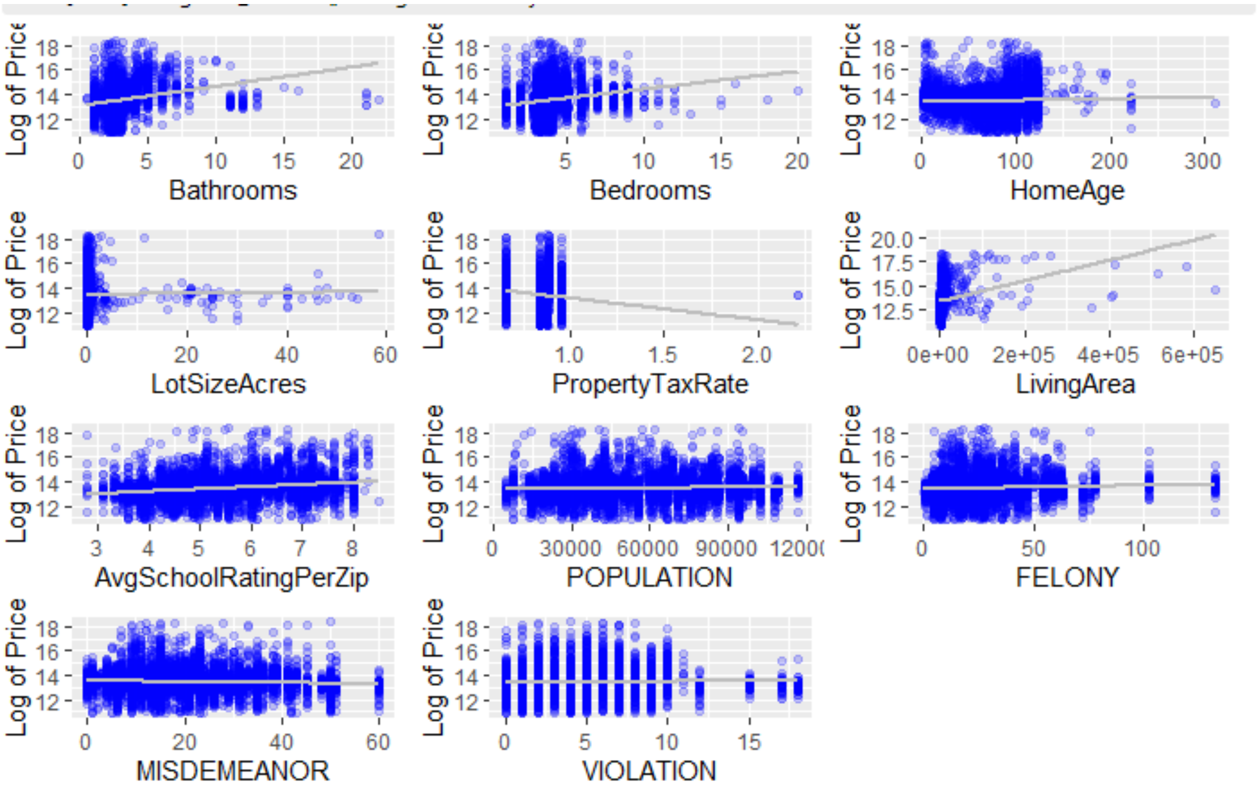


Figure 5: Scatter plots for log(Price Sold) vs. continuous predictors

Understanding price variability within zip codes can hint how complicated the model must be to represent the underlying patterns in the data. If there is a lot of variability within a zip code, it means that there are many other factors that are influencing price within the area that may not be represented in our dataset or that the model may need to include additional variables, making it complex and less interpretable. A plot of the standard deviation of sale prices by zip code is shown below in Figure 6:

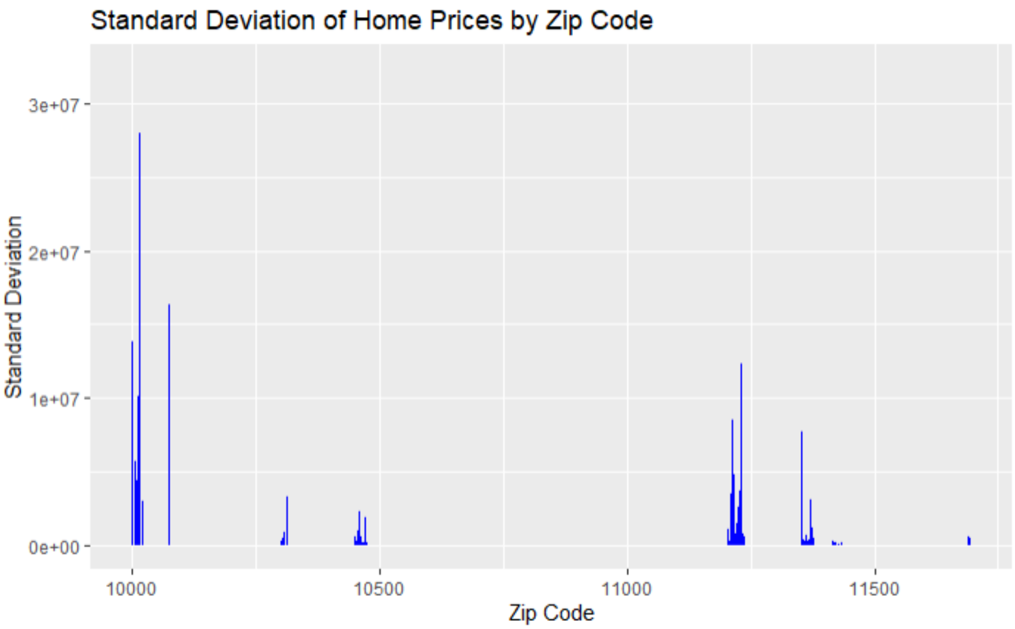


Figure 6: Plot of standard deviation of Price Sold vs. zip codes

Zip codes 10031 and 10016 show the most significant variability within the dataset, with a standard deviation of $32.37 and $27.96 million respectively.

Data Modeling

The data was split into a training, validation, and testing data set. The training set was obtained through random selection of 60% of the data, half of the remaining 40% was randomly selected to be the validation set, and the remaining data became the testing set. Linear regression was run using all predictors to see what R-squared value can be expected from the model. Variables were then removed based on subsequent transformations. A box-cox log transformation was applied to the dependent variable (PriceSold) to increase the normality of the distribution and scaled the price to the other predictors in Model 2 below.

Table 1: Different Preliminary Linear Regression models and their (Adjusted) R^2 values

Model	Model Formula	R^2	Adjusted R^2
0	lm(PriceSold ~ . -ZIPCODE - log_PriceSold, data=train)	0.1193	0.1177
1	lm(log_PriceSold ~ . -PriceSold - ZIPCODE, data=train)	0.2908	0.2895
2	lm(bc_PriceSold ~ . -ZIPCODE - log_PriceSold - PriceSold, data=train) Note that bc_PriceSold is the box-cox transformation of PriceSold	0.1936	0.1921
3	lm(PriceSold ~ . -ZIPCODE + factor(ZIPCODE) - log_PriceSold, data=train)	0.2717	0.2605
4	lm(log_PriceSold ~ . -ZIPCODE + factor(ZIPCODE) - PriceSold, data=train)	0.4833	0.4753
5	lm(log_PriceSold ~ . -HasFinishedBasement -HasSpa -HasLotSize -VIOLATION -PriceSold -ZIPCODE + factor(ZIPCODE), data=train) Remove predictors with p-values >= 0.5 from Model 4	0.4832	0.4754

6	lm(log_PriceSold ~ . -ZIPCODE + factor(ZIPCODE) -PriceSold -MISDEMEANOR -VIOLATION -POPULATION, data=train) Remove predictors that have correlation >= 0.5 from Model 4	0.4833	0.4753
7	lm(log_PriceSold ~ . -ZIPCODE + factor(ZIPCODE) - PriceSold -HasBasement -HasFinishedBasement + Basement, train) Created a 3 factor predictor called 'Basement' that captures 'NoBasement,' 'Unfinished,' or 'Finished'	0.4833	0.4753
8	lm(log_PriceSold ~ . -ZIPCODE + factor(ZIPCODE) - PriceSold, data=updated_train) updated_train here does a log transformation of 'HomeAge,' 'LivingArea,' and 'LotSizeAcres'	0.5349	0.5277

As can be seen in Table 1 above, the box-cox transformation (Model 2) yielded an improved R-squared value over the original (Model 0) but worse than the log transformed PriceSold (Model 1). Models 3 and 4 show the impact of adding 'ZIPCODE' as a factor predictor into the mix and it improved the R-squared value drastically in the case of log(PriceSold) in Model 4. Models 5-7 (all using model 4 as a base) showed the results when we removed predictors with p-values >=0.5, predictors with correlations >= 0.5, or when we created a 3-factor predictor 'Basement' to indicate whether the basement is finished, unfinished, or there isn't a basement. These models did not really change the R-squared value. Surprisingly, some predictors are actually not strongly correlated, such as living area and bedrooms (they have a correlation value of 0.02 which indicates that these two predictors are strongly independent). Correlations between all predictors can be seen in the Appendix under the "Other Visuals" section. Finally, model 8 runs using a log transformation of 'HomeAge,' 'LivingArea,' and 'LotSizeAcres' on top of base model 4. Based on the adjusted R-squared values, model 8 has the highest, but we will investigate model 4 in the next sections as our base case without transformations to the independent variables first.

Principal Component Regression (PCR)

Principal component regression was used to see if we can increase the R-squared value in our models. PCR, using the PLS package in R, was run on five different models: the first one is with some variables removed, the second one is with all variables, third one is with log transformation for "PriceSold", the fourth one is log transformation on PriceSold, HomeAge, LivngArea, and LotSizeAcres, and the last one is including every factor in the fourth one, but adding ZIPCODE as a factor variable.

Looking at the cross-validation error for the first PCR model, the smallest error occurs at 15 components. However, cross-validation error is almost the same after 6 components. In this case, using the smallest number of components would be suffice, so for the first PCR model, 6 components would be suitable. For the second PCR model, we can see that 8 components have the lowest error, and it increases after that, so 8 components would be suitable. The third and fourth PCR model, 11 components would be suitable since there are no significant changes after that. The fifth PCR model uses zip code as a factor. The fifth model shows dramatic decrease of cross-validation error until 16 components. After that, it continuously decreases, and the smallest error occurs at 178 components.

Comparing the R-squared values with the linear regression models, most of the PCR models performed worse nor showed any significant improvement on models. For the fifth PCR model, it shows that there is high R-adjusted value, however, this is likely overfitted due to adding zip code as a factor variable and having limited data per zip code, so we will reject this model.

Feature Engineering

Three methods were used for feature selection: decision tree, stepwise selection, and Lasso regression. The decision tree was trained on two versions of input data. The first version used zip code as a factor variable, and the second version excluded zip code. Although the first tree model had better performance on the validation data set, it had too many variables and was hard to interpret. Thus, the team decided to move forward with the simpler tree model. This tree was first split on living area and then on school rating, felony, bathroom, and home age. A screenshot of the tree was included in the Appendix.

Stepwise selection was also used. The starting model used log of price as the dependent variable, and used all other variables, including zip codes, as independent variables. The final model after performing stepwise selection was included in the Appendix. The final model had 95 variables, where 81 out of 95 were zip codes. This is in line with our expectation, where selected areas of New York may not follow the generic pricing model. For example, the coefficient of zip code 10021 shows that homes in Upper East Manhattan have higher values than equivalent homes elsewhere. In addition, we are not surprised to see that factors such as living area, bathrooms, and bedrooms are positively correlated with sales price. Fireplace, cooling system, and strong school rating are also expected to increase home prices. On the other hand, HOA fee, violations, and misdemeanor negatively impact home prices. We do want to call out that the high volume of zip code variables (161 of them) indicates a risk of overfitting (limited observations per zip code). If we had more time in the project, we would try to group the zip codes into larger areas to reduce the number of variables.

Two Lasso models were used. One has the penalty term based on lambda.min and the other is based on lambda.1se. The model with lambda.1se has fewer features but performed worse on the validation data set. The key features in this model are: “PropertyTaxRate,” “LivingArea,” “Bathrooms,” “Bedrooms,” “HasFireplace,” “AvgSchoolRatingPerZip”, and “FELONY.”

Model Comparison and Performance

Model validation is based on performance on the validation dataset, which is 20% of the data. The model performance was measured based on sum of squared error on log price. Here are the results in Table 2:

Table 2: Model Performance based on SSE

Model	**Regression tree, with zipcode**	**Regression tree, without zipcode**	**Step wise selection**	**Lasso, lambda min**	**Lasso, lambda 1se**
Sum of squared error (log price)	1038.768	1124.468	921.5737	1241.06	1276.446

Note that these numbers will vary depending on the training/validation split. We recommend moving forward with the stepwise selection model since it has the best performance. This model has an adjusted R-squared of 0.5097.

In addition, Cook’s Distance can be applied to this chosen model to identify influential observations that may have a large impact on the estimated regression and can affect the overall fit of the model. A common threshold for Cooks Distance is between 0.1 and 1. A lower threshold such as 0.1 might be appropriate for this model to account for the increased complexity of the number of predictors considered. However, the data set is large and so can tolerate a higher number of influential observations without affecting the fit of the model so the Cooks threshold may not need to be so low and can be decided based on the model. Cooks distance graph for this selected model can be seen in the Appendix under “Other Visuals” section. Note that we do not see high values of Cooks distance.

Checking Model Assumptions

Plotting the standardized residuals versus the predictors for the multiple linear regression model using predictors selected from stepwise model in Figure 7 allows us to check the linearity assumption. The linearity assumption states that the changes in the independent variables are associated with a constant change in the dependent variable.

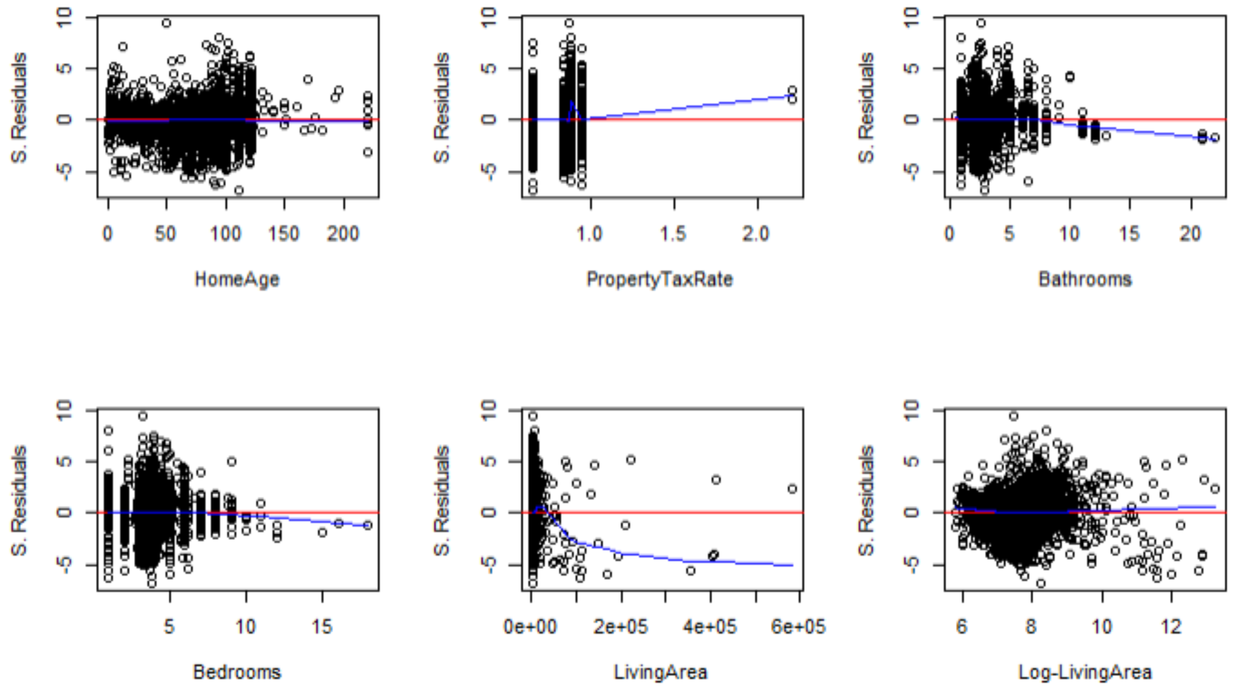


Figure 7: Standardized residuals vs different predictors for stepwise model (final model)

For the predictors chosen from the stepwise feature selection, we can see that the residuals and the smooth fit blue line exhibit a linear shape and so the linearity assumption appears to be satisfied. Living area appears to exhibit some non-linearity and heteroskedasticity but transforming it by taking the log restores the constant variance and linearity assumptions.

For the plot of the standardized residuals vs fitted values in Figure 8, we see a similar linear pattern which confirms that the linearity assumption holds. However, the plot also shows that the spread of residuals is roughly equal per fitted value, with the variance increasing as fitted value increases. This appears to meet the assumption that the variance of the errors is constant for all values of the independent variables. There also appears to be some clustering.

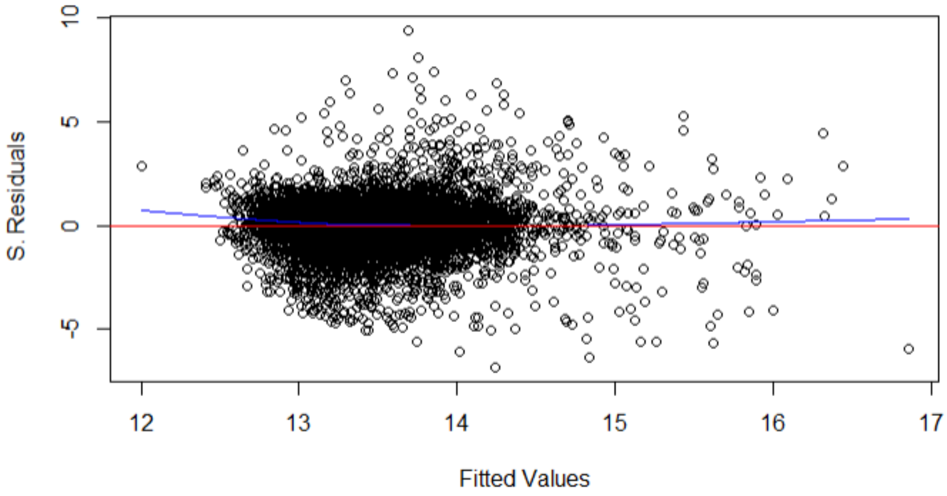


Figure 8: Standardized residuals vs Fitted Values for stepwise model (final model)

In Figure 9, from the histogram, we see a normal peak and a good spread of residuals. However, the QQ plot reveals heavy left and right tails, suggesting that the residuals are not normally distributed and a violation of the normality assumption.

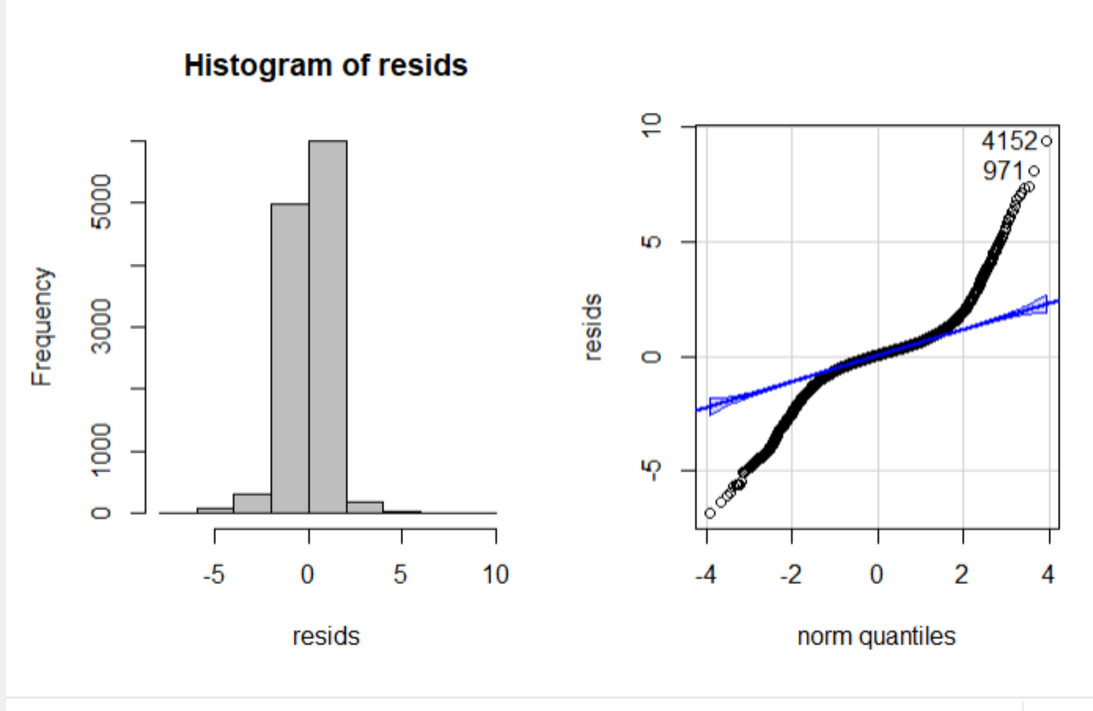


Figure 9: Histogram and QQ Plot for stepwise model (final model)

Discussion

The final model did not have as good of a normal distribution as we would have liked, as shown in Figure 9 above. We tried several models and the results were similar.

As expected, living area, lot size, bathrooms, bedrooms, average school ratings per zip code, and having a fireplace all have a positive impact on housing price. Of these predictors, living area has the most positive coefficient.

Surprisingly, the model also showed a positive impact on housing price for the number of felonies in a neighborhood, which seemed counterintuitive. This could be due to the limited data present per zip code in our model that could possibly skew the results. With over 80 zip codes used in the final model, the distribution of data across these zip codes are unequal and may not be sufficient, thus yielded a positive correlation to price. In addition, some zip codes displayed a positive correlation to housing price due to the location within NYC.

Predictors that have a negative impact on housing price include having an HOA, having misdemeanors/violations in the neighborhood, property tax rate, population, and home age. Population being negatively correlated to housing price also seems surprising since more densely populated areas like Manhattan are usually more expensive, although the coefficient is very small relative to the other predictors. Again, some zip codes also have a negative correlation to housing price due to their location within NYC.

Conclusion

Research Question 1: What are the best predictors for NYC housing prices? The predictors can be internal (such as square footage, number of bedrooms) or external (such as neighborhood crime rate).

For both the initial model with all predictors and the best performing optimized model using only the selected predictors, the predictors with the highest coefficients and therefore the ones with the highest effect on price are the zip code in which the home is located. The 4 zip codes with the highest positive influence on price are 10011, 10021, 10023, and 10036. These are all in Manhattan, an area famous for its significantly higher than average living costs, and so it makes sense that these zip codes would be a predictor of high housing prices. Other influential predictors were property tax rate (B = -1.161), if the property has an HOA (B = -0.2286), and if the property has a fireplace (B = 0.1431). Since this is a Log-Linear model for these predictors, price changes by 100*(coefficient)% for a one unit increase in the independent variable with all other variables in the model held constant. For example, the model predicts that having a fireplace would increase the sale price by 14.31%, having an HOA can will decrease it by 22.86%, and being in zip code 10036 will increase price by 282.1%. Overall, the model predicts that various zip codes have the largest effect on housing prices.

Research Question 2: Which neighborhoods have the highest variation of property prices?

The five zip codes with the highest variability in prices are 10031, 10016, 10036, 10005, and 11101. These largely overlap with the most affluent and influential zip codes in the model. One hypothesis is that there is a mixture of old and new housing, leading to a mix of affordable and high-end housing options that contributes to the variability in home prices within that zip code. For example, 10031 is in Harlem, which has undergone significant gentrification in the past few years. Another potential reason is that more expensive neighborhood will naturally have a higher variation in pricing since the overall sales prices are higher and homes are not as easily appraised nor have a set market value as conventional homes.

Research Question 3: Given a single active listing, does the listing price represent the true property value or is it intentionally set low to trigger a bidding war?

We can use the proposed model to predict a property's value. We identified two methods to determine whether a given home's price is intentionally set low to trigger a bidding war. The first method is to compare the predicted sales price to the listing prices of homes that are for sale (this was a group of data we saved from earlier). However, since our model is not able to explain the full price variation, we would like to add a buffer to the predicted price. If the predicted price is 30% higher than the listed price, we should expect a bidding war to happen. Alternatively, we can use the upper predicted interval. If the upper interval is more than 250% higher than the listed price, we should expect a bidding war. These thresholds are chosen judgmentally. Please see Figure 10 below for the areas that we deem indicate bidding wars in each plot. We would gather more data to validate these thresholds if we had more time for the project.

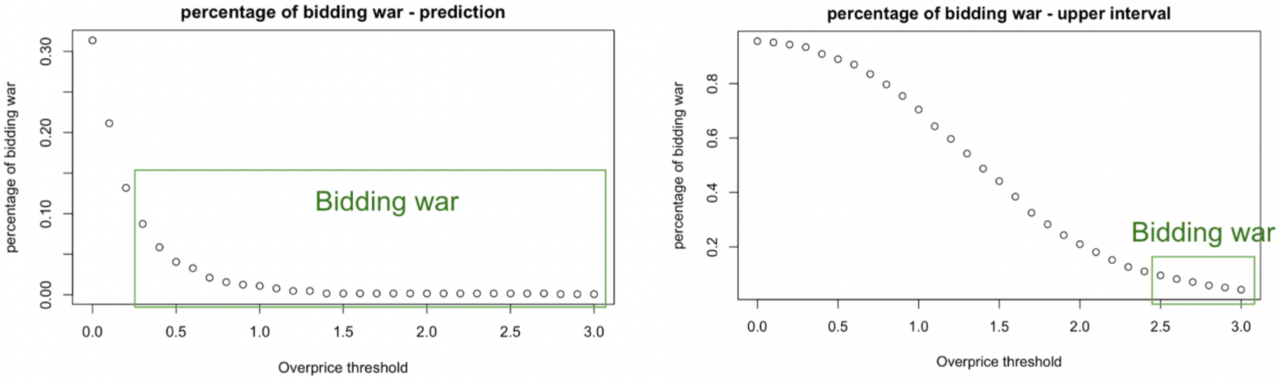


Figure 10: Percentage of homes listed for bidding war vs. different Overprice Thresholds

Further Research

The current model has two limitations. The first limitation is that our input data did not factor in macro-economic changes. Our input data is limited to 2021. The assumption is that the period is short enough that macroeconomic factors have a uniform impact on all homes. However, 2021 was a year full of macroeconomic changes. The inflation rate kept rising, the fiscal policy became tighter at the end of the year, and energy and commodity prices also had some fluctuation throughout the year. Given that home prices can be impacted by the macroeconomic factors (RBC, 2013), we will incorporate those financial factors into our model if we have more time. For example, we may be able to tie home prices to inflation rates or find correlations between home prices and prices of alternative investments. However, these impacts are more difficult to model. Some financial factors can have delayed impact on home prices, and some can have more severe impacts on selected areas of the city. Therefore, a much more complicated model will be required to factor in the macroeconomic impacts.

The second limitation in our model is that we did not consider seasonality. Usually, more real estate transactions happen in the summer than in the winter. Therefore, it is reasonable to assume that seasonality has some impact on home prices. If we have more time with the project, we will train the model on multiple years of data and will remove trend and seasonality before feeding the data into the regression model.

References

- Ceccat, V. C., & Wilhelmsson, M. W. (2011). THE IMPACT OF CRIME ON APARTMENT PRICES: EVIDENCE FROM STOCKHOLM, SWEDEN. Geografiska Annaler. Series B, Human Geography, 91(1). https://Stable URL: https://www.jstor.org/stable/41315194
- RBC. (2013, November). Priced out: Understanding the factors affecting home prices in ... - RBC. Retrieved March 21, 2023, from http://www.rbc.com/community-sustainability/_assets-custom/pdf/Priced-Out-RBC-Pembina.pdf
- The Colorado College, & Ockey, T. (2018). HOUSE PRICES: FACTORS THAT INFLUENCE THE VALUE OF A HOME

Appendix

The following contains supplementary information, chart/graphs, links, and descriptions to our project.

Predictor Descriptions

Table 3: Predictors and their descriptions/units in cleaned data set

Variable	Description/Units	Variable	Description/Units
ZIPCODE	5-digit zip code	HasHeating	Binary
City	String text of neighborhood	HasCooling	Binary
HomeStatus	Sold / recently sold/ for sale	HasPrivatePool	Binary
PriceSold	Unit: dollar	HasSpa	Binary
HomeAge	Age of home from year 2021	HasHOA	Binary
PropertyTaxRate	Tax rate of home	HasLotSize	Binary
LivingArea	Size of living area in sq. ft.	LotSizeAcres	Size of lot in acres
Bathrooms	Number of bathrooms	AvgSchoolRatingPerZip	Average rating of elementary, middle, and high school
Bedrooms	Number of bedrooms	Population	Population size
HasBasement	Binary	FELONY	Number of reported felonies in zip code
HasFinishedBasement	Binary	MISDEMEANOR	Number of reported misdemeanors in zip code
HasGarage	Binary	VIOLATION	Number of reported crimes in zip code that are not felonies or misdemeanors
HasFireplace	Binary		

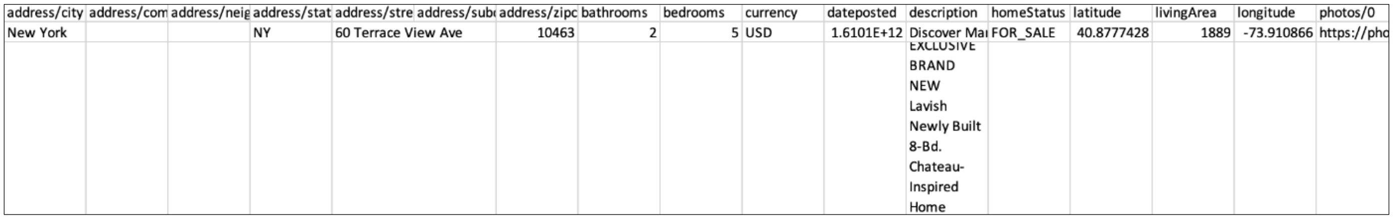
Data Sets

Links to the original data sets as well as the merged/cleaned data set are listed in Table 4 below.

Table 4: Links to different data sets

Original Zillow Data Set	Original Population Data Set	Original Crime Data Set	Merged and Cleaned Data Set

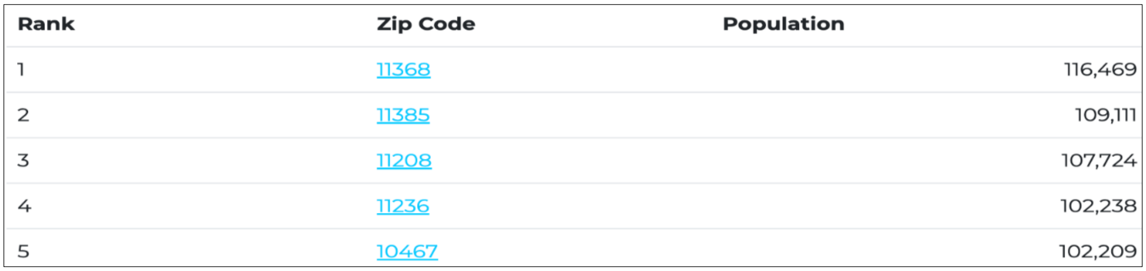
Figure 11 shows a screenshot of some of the original columns of the Zillow data set. Note that there are a lot of predictors that can be skipped because they are not relevant or are not needed for the purpose of doing linear regression. We went thru all 1,507 predictors to get to the 25 shown in this report.



address/city	address/com	address/neig	address/stat	address/stre	address/sub	address/zipc	bathrooms	bedrooms	currency	dateposted	description	homeStatus	latitude	livingArea	longitude	photos/0
New York			NY	60 Terrace View Ave		10463	2	5	USD	1.6101E+12	Discover Mai EXCLUSIVE BRAND NEW Lavish Newly Built 8-Bd. Chateau- Inspired Home	FOR_SALE	40.8777428	1889	-73.910866	https://pho

Figure 11: Screenshot of some of the original columns of Zillow data set

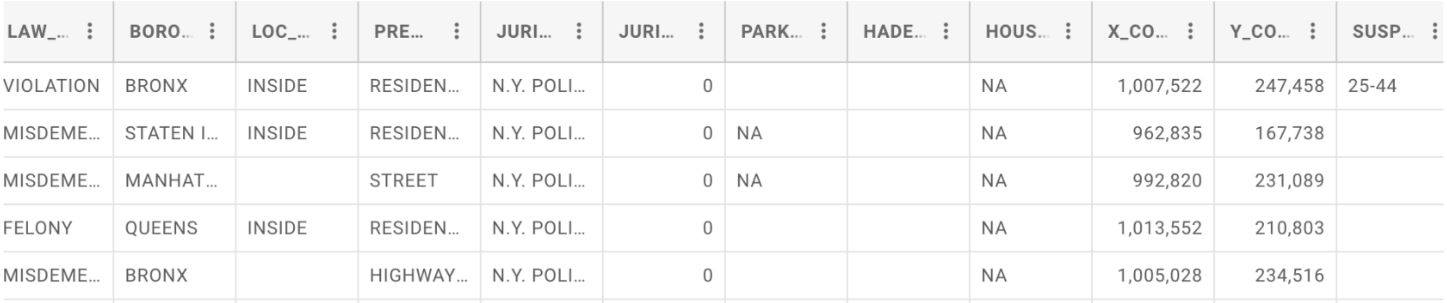
Figure 12 shows a screenshot of the population data from the American Community Survey (ACS). Note that this was in a table format on the website, and we had to convert to a csv file to merge with the Zillow data set.



Rank	Zip Code	Population
1	11368	116,469
2	11385	109,111
3	11208	107,724
4	11236	102,238
5	10467	102,209

Figure 12: Screenshot of 2021 population data from ACS

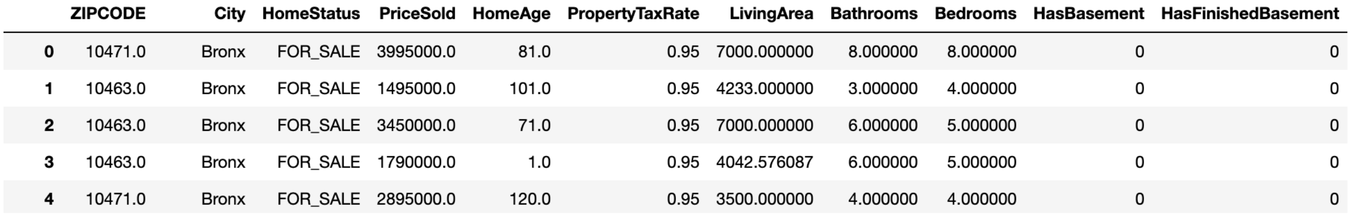
Figure 13 shows a screenshot of part of the crime data set. Note that there are many columns, including the latitude and longitude variables not shown here. The zip code was retrieved using latitude and longitude via the GeoPy API call. The number of felonies, violations, and misdemeanors were counted from the first column shown below and split into 3 different variables in the cleaned data set.



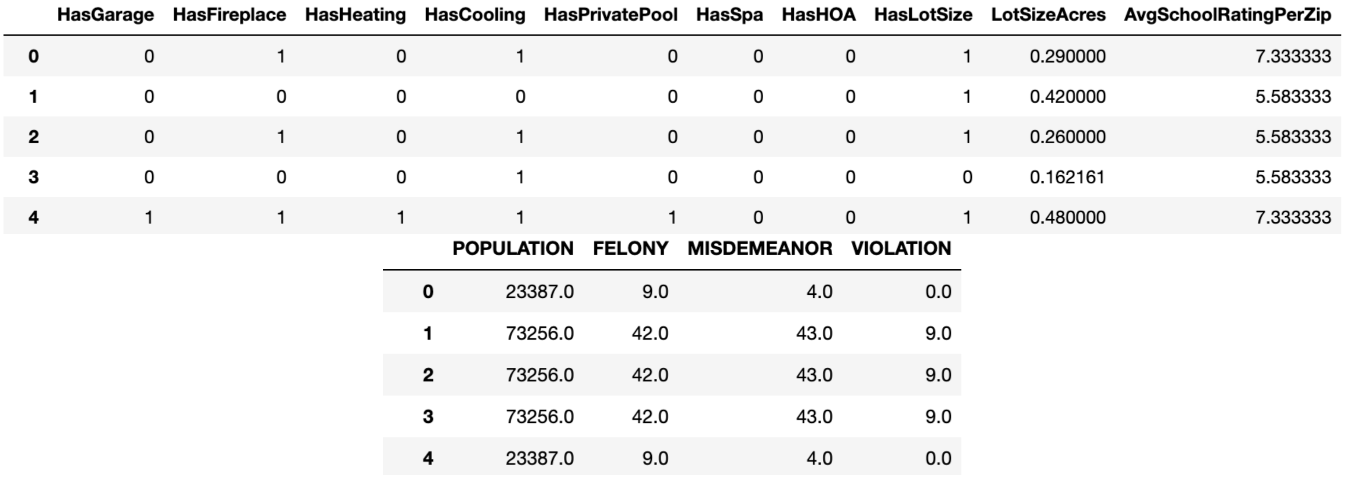
LAW_…	BORO…	LOC_…	PRE…	JURI…	JURI…	PARK…	HADE…	HOUS…	X_CO…	Y_CO…	SUSP…
VIOLATION	BRONX	INSIDE	RESIDEN…	N.Y. POLI…	0			NA	1,007,522	247,458	25-44
MISDEME…	STATEN I…	INSIDE	RESIDEN…	N.Y. POLI…	0	NA		NA	962,835	167,738	
MISDEME…	MANHAT…		STREET	N.Y. POLI…	0	NA		NA	992,820	231,089	
FELONY	QUEENS	INSIDE	RESIDEN…	N.Y. POLI…	0			NA	1,013,552	210,803	
MISDEME…	BRONX		HIGHWAY…	N.Y. POLI…	0			NA	1,005,028	234,516	

Figure 13: Screenshot of some columns of the original NYPD crime data set

Figure 14 shows screenshots of the full 25 predictors and a few rows of the cleaned data set, including the 'FELONY', 'MISDEMEANOR', and 'VIOLATION' variable counts mentioned above.



	ZIPCODE	City	HomeStatus	PriceSold	HomeAge	PropertyTaxRate	LivingArea	Bathrooms	Bedrooms	HasBasement	HasFinishedBasement
0	10471.0	Bronx	FOR_SALE	3995000.0	81.0	0.95	7000.000000	8.000000	8.000000	0	0
1	10463.0	Bronx	FOR_SALE	1495000.0	101.0	0.95	4233.000000	3.000000	4.000000	0	0
2	10463.0	Bronx	FOR_SALE	3450000.0	71.0	0.95	7000.000000	6.000000	5.000000	0	0
3	10463.0	Bronx	FOR_SALE	1790000.0	1.0	0.95	4042.576087	6.000000	5.000000	0	0
4	10471.0	Bronx	FOR_SALE	2895000.0	120.0	0.95	3500.000000	4.000000	4.000000	0	0



	HasGarage	HasFireplace	HasHeating	HasCooling	HasPrivatePool	HasSpa	HasHOA	HasLotSize	LotSizeAcres	AvgSchoolRatingPerZip
0	0	1	0	1	0	0	0	1	0.290000	7.333333
1	0	0	0	0	0	0	0	1	0.420000	5.583333
2	0	1	0	1	0	0	0	1	0.260000	5.583333
3	0	0	0	1	0	0	0	0	0.162161	5.583333
4	1	1	1	1	1	0	0	1	0.480000	7.333333

	POPULATION	FELONY	MISDEMEANOR	VIOLATION
0	23387.0	9.0	4.0	0.0
1	73256.0	42.0	43.0	9.0
2	73256.0	42.0	43.0	9.0
3	73256.0	42.0	43.0	9.0
4	23387.0	9.0	4.0	0.0

Figure 14: Screenshots of the full 25 predictors of the cleaned data set

Steps Taken For Data Clean Up

The following steps were taken to clean up the merged data set:

- Filter data set to only have homes in NY, a home status of ‘SOLD,’ ‘RECENTLY_SOLD,’ or ‘FOR_SALE,’ and home type to be single family
- Average all school ratings within a zip code, create a new predictor “AvgSchoolRatingPerZip,” and merge into data set by zip code
- Removed predictors that did not have data for the majority of rows or are repeats of previous predictors ('TotalBath,' 'FullBath,' 'HalfBath,' 'Style,' 'HasAttachedGarage,' 'GarageSpaces,' 'HasOpenParking,' 'Parking,' 'HasAttachedProperty,' 'HasView,' 'IsNewConstruction,' 'HasCarport')
- Created 2 new predictors (‘HasBasement‘ and ‘HasFinishedBasement’) based off of the string values in ‘BasementFinish’ predictor to determine if a home has a basement and if it is finished
- Created a new predictor ‘HasHOA’ to determine if a home has an HOA from a string variable ‘LotSizeHOA’
- Created a new predictor ‘HOA’ to record the HOA amount from a mixed alphanumeric variable ‘LotsizeHOAValue’
- Created 2 new predictors from the alphanumeric variable ‘LotSize’: ‘HasLotSize’ to determine if a home has a lot size area available and ‘LotSizeAcres’ to convert available lot sizes to standard units of acres
- Create a new predictor ‘HomeAge’ from variable ‘YearBuilt’ to indicate age of home
- Convert all binary variables from True/False to 1/0
- Replace ‘NaN’ with 0 for predictors ‘POPULATION,’ ‘FELONY,’ ‘MISDEMEANOR,’ and ‘VIOLATION’
- Remove rows where either ‘ZIPCODE,’ ‘AvgSchoolRatingPerZip,’ or ‘PropertyTaxRate ’ = ‘NaN’
- Remove rows where ‘POPULATION’ is 0
- Drop continuous ‘HOA’ predictor because it’s hard to tell is considered a valid HOA fee, but keep binary predictor ‘HasHOA’
- Impute ‘HomeAge’ for rows without a home age using the mean home age for that zip code
- Impute ‘LotSizeAcres’ for rows without a lot size using the mean lot size for that zip code
- Impute ‘PriceSold,’ our dependent variable, for rows without a price sold using the mean price sold for that zip code. Also based on research, it is unlikely that a home sells for less than $50,000, so we also replaced ‘PriceSold’ < $50,000 with the imputed mean price sold by zip code.
- Impute ‘LivingArea’ for rows without a living area size using the mean living area for that zip code and for rows where ‘LivingArea’ <= 300 sq ft, since that is unlikely.
- Remove rows where either ‘Bathrooms’ or ‘Bedrooms’ is greater than 25, since that is unlikely. Then, impute ‘Bathrooms’ and ‘Bedrooms’ for rows without number of bathrooms or bedrooms, respectively, by using the

mean bathroom or bedroom for that zip code. Finally, remove rows where either 'Bathrooms' or 'Bedrooms' is still null.

- Remove duplicated rows

Return to "Data Cleaning" section.

Other Visuals

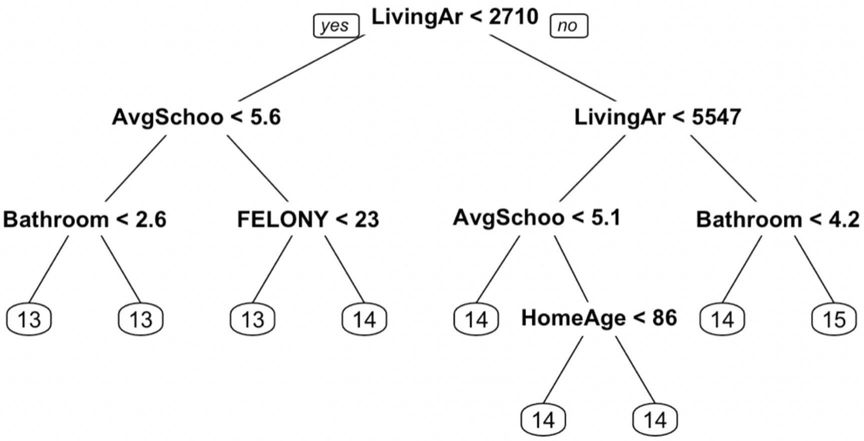
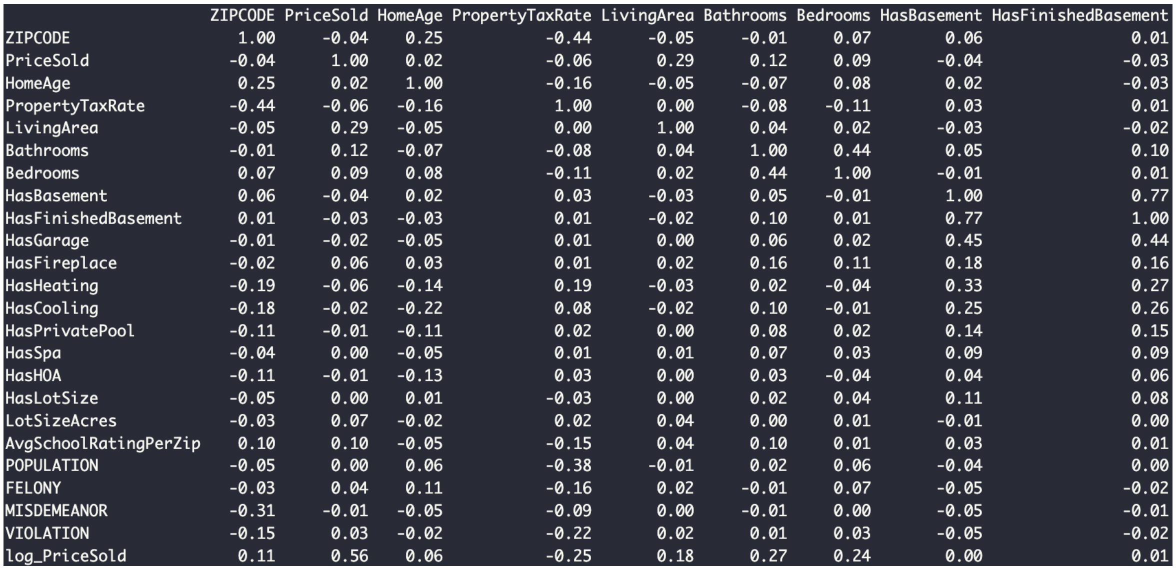
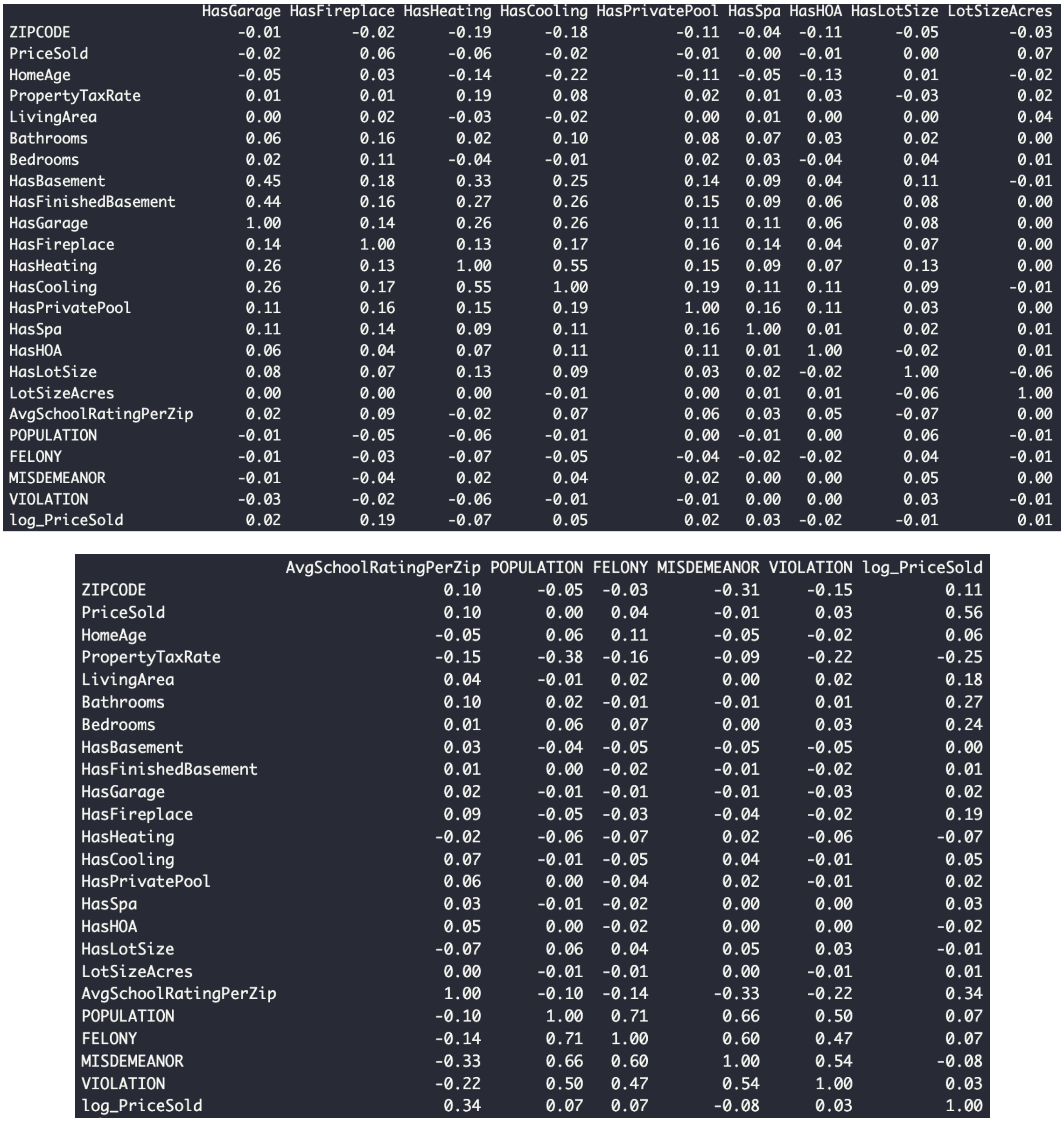


Figure 15: Decision Tree model



```
                      ZIPCODE PriceSold HomeAge PropertyTaxRate LivingArea Bathrooms Bedrooms HasBasement HasFinishedBasement
ZIPCODE                  1.00     -0.04    0.25           -0.44      -0.05     -0.01     0.07        0.06                0.01
PriceSold               -0.04      1.00    0.02           -0.06       0.29      0.12     0.09       -0.04               -0.03
HomeAge                  0.25      0.02    1.00           -0.16      -0.05     -0.07     0.08        0.02               -0.03
PropertyTaxRate         -0.44     -0.06   -0.16            1.00       0.00     -0.08    -0.11        0.03                0.01
LivingArea              -0.05      0.29   -0.05            0.00       1.00      0.04     0.02       -0.03               -0.02
Bathrooms               -0.01      0.12   -0.07           -0.08       0.04      1.00     0.44        0.05                0.10
Bedrooms                 0.07      0.09    0.08           -0.11       0.02      0.44     1.00       -0.01                0.01
HasBasement              0.06     -0.04    0.02            0.03      -0.03      0.05    -0.01        1.00                0.77
HasFinishedBasement      0.01     -0.03   -0.03            0.01      -0.02      0.10     0.01        0.77                1.00
HasGarage               -0.01     -0.02   -0.05            0.01       0.00      0.06     0.02        0.45                0.44
HasFireplace            -0.02      0.06    0.03            0.01       0.02      0.16     0.11        0.18                0.16
HasHeating              -0.19     -0.06   -0.14            0.19      -0.03      0.02    -0.04        0.33                0.27
HasCooling              -0.18     -0.02   -0.22            0.08      -0.02      0.10    -0.01        0.25                0.26
HasPrivatePool          -0.11     -0.01   -0.11            0.02       0.00      0.08     0.02        0.14                0.15
HasSpa                  -0.04      0.00   -0.05            0.01       0.01      0.07     0.03        0.09                0.09
HasHOA                  -0.11     -0.01   -0.13            0.03       0.00      0.03    -0.04        0.04                0.06
HasLotSize              -0.05      0.00    0.01           -0.03       0.00      0.02     0.04        0.11                0.08
LotSizeAcres            -0.03      0.07   -0.02            0.02       0.04      0.00     0.01       -0.01                0.00
AvgSchoolRatingPerZip    0.10      0.10   -0.05           -0.15       0.04      0.10     0.01        0.03                0.01
POPULATION              -0.05      0.00    0.06           -0.38      -0.01      0.02     0.06       -0.04                0.00
FELONY                  -0.03      0.04    0.11           -0.16       0.02     -0.01     0.07       -0.05               -0.02
MISDEMEANOR             -0.31     -0.01   -0.05           -0.09       0.00     -0.01     0.00       -0.05               -0.01
VIOLATION               -0.15      0.03   -0.02           -0.22       0.02      0.01     0.03       -0.05               -0.02
log_PriceSold            0.11      0.56    0.06           -0.25       0.18      0.27     0.24        0.00                0.01
```



```
                      HasGarage HasFireplace HasHeating HasCooling HasPrivatePool HasSpa HasHOA HasLotSize LotSizeAcres
ZIPCODE                   -0.01        -0.02      -0.19      -0.18          -0.11  -0.04  -0.11      -0.05        -0.03
PriceSold                 -0.02         0.06      -0.06      -0.02          -0.01   0.00  -0.01       0.00         0.07
HomeAge                   -0.05         0.03      -0.14      -0.22          -0.11  -0.05  -0.13       0.01        -0.02
PropertyTaxRate            0.01         0.01       0.19       0.08           0.02   0.01   0.03      -0.03         0.02
LivingArea                 0.00         0.02      -0.03      -0.02           0.00   0.01   0.00       0.00         0.04
Bathrooms                  0.06         0.16       0.02       0.10           0.08   0.07   0.03       0.02         0.00
Bedrooms                   0.02         0.11      -0.04      -0.01           0.02   0.03  -0.04       0.04         0.01
HasBasement                0.45         0.18       0.33       0.25           0.14   0.09   0.04       0.11        -0.01
HasFinishedBasement        0.44         0.16       0.27       0.26           0.15   0.09   0.06       0.08         0.00
HasGarage                  1.00         0.14       0.26       0.26           0.11   0.11   0.06       0.08         0.00
HasFireplace               0.14         1.00       0.13       0.17           0.16   0.14   0.04       0.07         0.00
HasHeating                 0.26         0.13       1.00       0.55           0.15   0.09   0.07       0.13         0.00
HasCooling                 0.26         0.17       0.55       1.00           0.19   0.11   0.11       0.09        -0.01
HasPrivatePool             0.11         0.16       0.15       0.19           1.00   0.16   0.11       0.03         0.00
HasSpa                     0.11         0.14       0.09       0.11           0.16   1.00   0.01       0.02         0.01
HasHOA                     0.06         0.04       0.07       0.11           0.11   0.01   1.00      -0.02         0.01
HasLotSize                 0.08         0.07       0.13       0.09           0.03   0.02  -0.02       1.00        -0.06
LotSizeAcres               0.00         0.00       0.00      -0.01           0.00   0.01   0.01      -0.06         1.00
AvgSchoolRatingPerZip      0.02         0.09      -0.02       0.07           0.06   0.03   0.05      -0.07         0.00
POPULATION                -0.01        -0.05      -0.06      -0.01           0.00  -0.01   0.00       0.06        -0.01
FELONY                    -0.01        -0.03      -0.07      -0.05          -0.04  -0.02  -0.02       0.04        -0.01
MISDEMEANOR               -0.01        -0.04       0.02       0.04           0.02   0.00   0.00       0.05         0.00
VIOLATION                 -0.03        -0.02      -0.06      -0.01          -0.01   0.00   0.00       0.03        -0.01
log_PriceSold              0.02         0.19      -0.07       0.05           0.02   0.03  -0.02      -0.01         0.01
```

```
                      AvgSchoolRatingPerZip POPULATION FELONY MISDEMEANOR VIOLATION log_PriceSold
ZIPCODE                                0.10      -0.05  -0.03       -0.31     -0.15          0.11
PriceSold                              0.10       0.00   0.04       -0.01      0.03          0.56
HomeAge                               -0.05       0.06   0.11       -0.05     -0.02          0.06
PropertyTaxRate                       -0.15      -0.38  -0.16       -0.09     -0.22         -0.25
LivingArea                             0.04      -0.01   0.02        0.00      0.02          0.18
Bathrooms                              0.10       0.02  -0.01       -0.01      0.01          0.27
Bedrooms                               0.01       0.06   0.07        0.00      0.03          0.24
HasBasement                            0.03      -0.04  -0.05       -0.05     -0.05          0.00
HasFinishedBasement                    0.01       0.00  -0.02       -0.01     -0.02          0.01
HasGarage                              0.02      -0.01  -0.01       -0.01     -0.03          0.02
HasFireplace                           0.09      -0.05  -0.03       -0.04     -0.02          0.19
HasHeating                            -0.02      -0.06  -0.07        0.02     -0.06         -0.07
HasCooling                             0.07      -0.01  -0.05        0.04     -0.01          0.05
HasPrivatePool                         0.06       0.00  -0.04        0.02     -0.01          0.02
HasSpa                                 0.03      -0.01  -0.02        0.00      0.00          0.03
HasHOA                                 0.05       0.00  -0.02        0.00      0.00         -0.02
HasLotSize                            -0.07       0.06   0.04        0.05      0.03         -0.01
LotSizeAcres                           0.00      -0.01  -0.01        0.00     -0.01          0.01
AvgSchoolRatingPerZip                  1.00      -0.10  -0.14       -0.33     -0.22          0.34
POPULATION                            -0.10       1.00   0.71        0.66      0.50          0.07
FELONY                                -0.14       0.71   1.00        0.60      0.47          0.07
MISDEMEANOR                           -0.33       0.66   0.60        1.00      0.54         -0.08
VIOLATION                             -0.22       0.50   0.47        0.54      1.00          0.03
log_PriceSold                          0.34       0.07   0.07       -0.08      0.03          1.00
```

Figure 16: Correlation Table for Untransformed Predictors

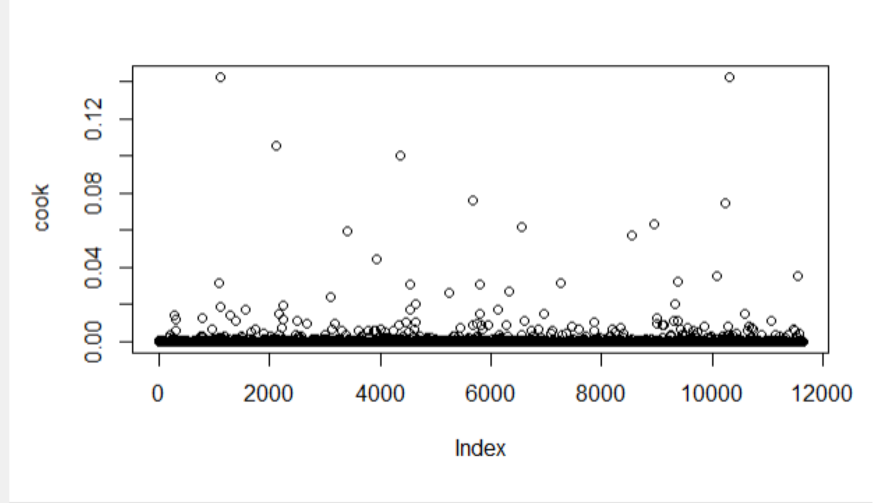
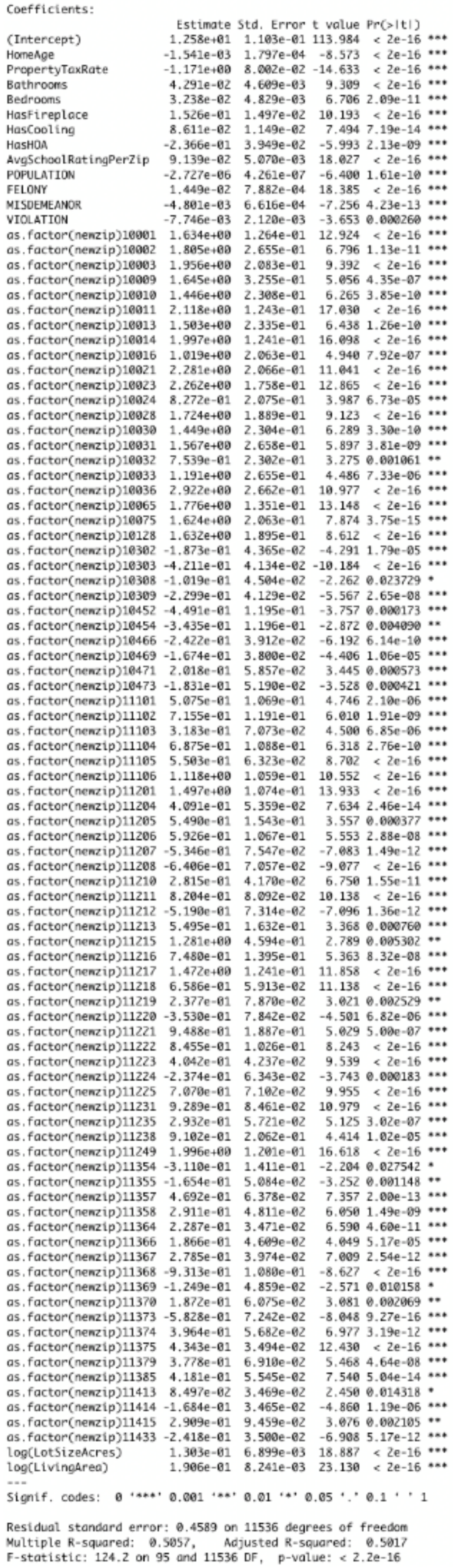


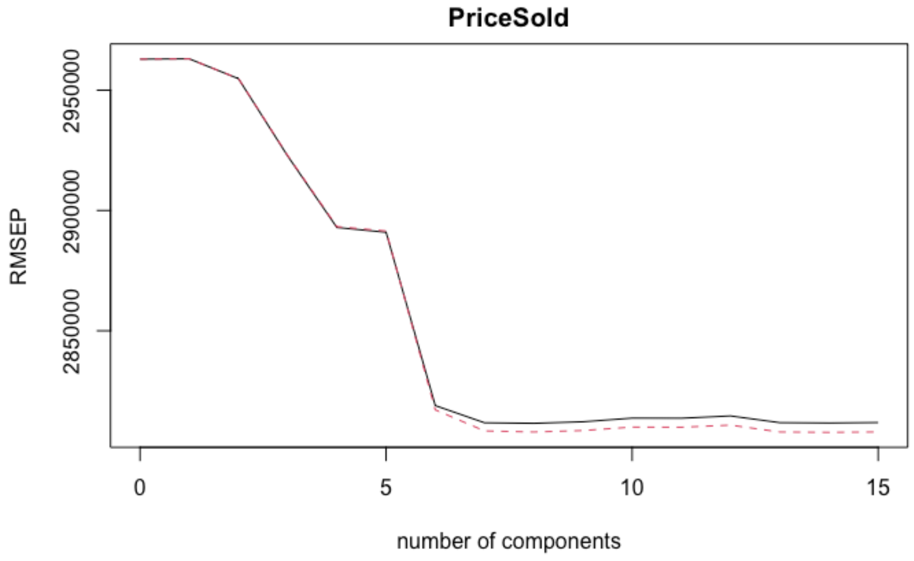
Figure 17: Cooks Distance for FInal Stepwise Model



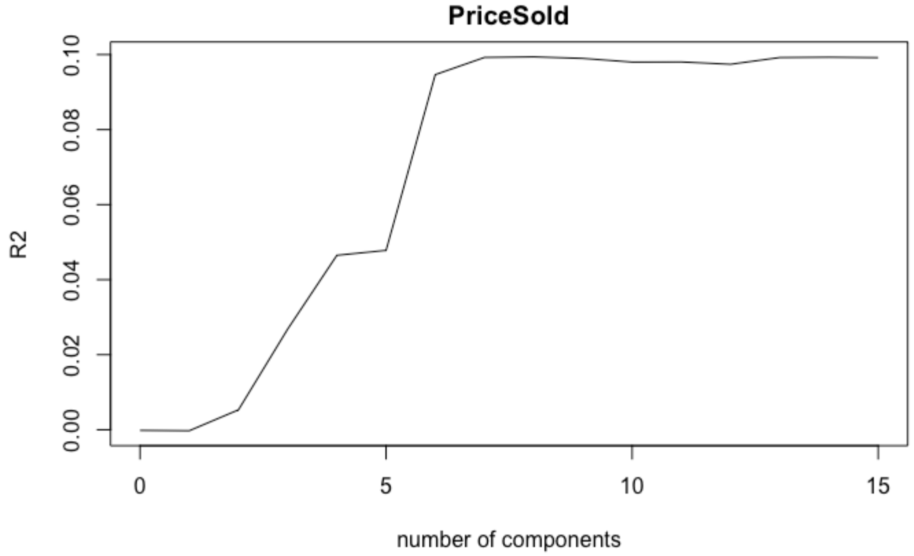
```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.258e+01  1.103e-01 113.984  < 2e-16 ***
HomeAge               -1.541e-03  1.797e-04  -8.573  < 2e-16 ***
PropertyTaxRate       -1.171e+00  8.002e-02 -14.633  < 2e-16 ***
Bathrooms              4.291e-02  4.609e-03   9.309  < 2e-16 ***
Bedrooms               3.238e-02  4.829e-03   6.706 2.09e-11 ***
HasFireplace           1.526e-01  1.497e-02  10.193  < 2e-16 ***
HasCooling             8.611e-02  1.149e-02   7.494 7.19e-14 ***
HasHOA                -2.366e-01  3.949e-02  -5.993 2.13e-09 ***
AvgSchoolRatingPerZip  9.139e-02  5.070e-03  18.027  < 2e-16 ***
POPULATION            -2.727e-06  4.261e-07  -6.400 1.61e-10 ***
FELONY                 1.449e-02  7.882e-04  18.385  < 2e-16 ***
MISDEMEANOR           -4.801e-03  6.616e-04  -7.256 4.23e-13 ***
VIOLATION             -7.746e-03  2.120e-03  -3.653 0.000260 ***
as.factor(newzip)10001  1.634e+00  1.264e-01  12.924  < 2e-16 ***
as.factor(newzip)10002  1.805e+00  2.655e-01   6.796 1.13e-11 ***
as.factor(newzip)10003  1.956e+00  2.083e-01   9.392  < 2e-16 ***
as.factor(newzip)10009  1.645e+00  3.255e-01   5.056 4.35e-07 ***
as.factor(newzip)10010  1.446e+00  2.308e-01   6.265 3.85e-10 ***
as.factor(newzip)10011  2.118e+00  1.243e-01  17.030  < 2e-16 ***
as.factor(newzip)10013  1.503e+00  2.335e-01   6.438 1.26e-10 ***
as.factor(newzip)10014  1.997e+00  1.241e-01  16.098  < 2e-16 ***
as.factor(newzip)10016  1.019e+00  2.063e-01   4.940 7.92e-07 ***
as.factor(newzip)10021  2.281e+00  2.066e-01  11.041  < 2e-16 ***
as.factor(newzip)10023  2.262e+00  1.758e-01  12.865  < 2e-16 ***
as.factor(newzip)10024  8.272e-01  2.075e-01   3.987 6.73e-05 ***
as.factor(newzip)10028  1.724e+00  1.889e-01   9.123  < 2e-16 ***
as.factor(newzip)10030  1.449e+00  2.304e-01   6.289 3.30e-10 ***
as.factor(newzip)10031  1.567e+00  2.658e-01   5.897 3.81e-09 ***
as.factor(newzip)10032  7.539e-01  2.302e-01   3.275 0.001061 **
as.factor(newzip)10033  1.191e+00  2.655e-01   4.486 7.33e-06 ***
as.factor(newzip)10036  2.922e+00  2.662e-01  10.977  < 2e-16 ***
as.factor(newzip)10065  1.776e+00  1.351e-01  13.148  < 2e-16 ***
as.factor(newzip)10075  1.624e+00  2.063e-01   7.874 3.75e-15 ***
as.factor(newzip)10128  1.632e+00  1.895e-01   8.612  < 2e-16 ***
as.factor(newzip)10302 -1.873e-01  4.365e-02  -4.291 1.79e-05 ***
as.factor(newzip)10303 -4.211e-01  4.134e-02 -10.184  < 2e-16 ***
as.factor(newzip)10308 -1.019e-01  4.504e-02  -2.262 0.023729 *
as.factor(newzip)10309 -2.299e-01  4.129e-02  -5.567 2.65e-08 ***
as.factor(newzip)10452 -4.491e-01  1.195e-01  -3.757 0.000173 ***
as.factor(newzip)10454 -3.435e-01  1.196e-01  -2.872 0.004090 **
as.factor(newzip)10466 -2.422e-01  3.912e-02  -6.192 6.14e-10 ***
as.factor(newzip)10469 -1.674e-01  3.800e-02  -4.406 1.06e-05 ***
as.factor(newzip)10471  2.018e-01  5.857e-02   3.445 0.000573 ***
as.factor(newzip)10473 -1.831e-01  5.190e-02  -3.528 0.000421 ***
as.factor(newzip)11101  5.075e-01  1.069e-01   4.746 2.10e-06 ***
as.factor(newzip)11102  7.155e-01  1.191e-01   6.010 1.91e-09 ***
as.factor(newzip)11103  3.183e-01  7.073e-02   4.500 6.85e-06 ***
as.factor(newzip)11104  6.875e-01  1.088e-01   6.318 2.76e-10 ***
as.factor(newzip)11105  5.503e-01  6.323e-02   8.702  < 2e-16 ***
as.factor(newzip)11106  1.118e+00  1.059e-01  10.552  < 2e-16 ***
as.factor(newzip)11201  1.497e+00  1.074e-01  13.933  < 2e-16 ***
as.factor(newzip)11204  4.091e-01  5.359e-02   7.634 2.46e-14 ***
as.factor(newzip)11205  5.490e-01  1.543e-01   3.557 0.000377 ***
as.factor(newzip)11206  5.926e-01  1.067e-01   5.553 2.88e-08 ***
as.factor(newzip)11207 -5.346e-01  7.547e-02  -7.083 1.49e-12 ***
as.factor(newzip)11208 -6.406e-01  7.057e-02  -9.077  < 2e-16 ***
as.factor(newzip)11210  2.815e-01  4.170e-02   6.750 1.55e-11 ***
as.factor(newzip)11211  8.204e-01  8.092e-02  10.138  < 2e-16 ***
as.factor(newzip)11212 -5.190e-01  7.314e-02  -7.096 1.36e-12 ***
as.factor(newzip)11213  5.495e-01  1.632e-01   3.368 0.000760 ***
as.factor(newzip)11215  1.281e+00  4.594e-01   2.789 0.005302 **
as.factor(newzip)11216  7.480e-01  1.395e-01   5.363 8.32e-08 ***
as.factor(newzip)11217  1.472e+00  1.241e-01  11.858  < 2e-16 ***
as.factor(newzip)11218  6.586e-01  5.913e-02  11.138  < 2e-16 ***
as.factor(newzip)11219  2.377e-01  7.870e-02   3.021 0.002529 **
as.factor(newzip)11220 -3.530e-01  7.842e-02  -4.501 6.82e-06 ***
as.factor(newzip)11221  9.488e-01  1.887e-01   5.029 5.00e-07 ***
as.factor(newzip)11222  8.455e-01  1.026e-01   8.243  < 2e-16 ***
as.factor(newzip)11223  4.042e-01  4.237e-02   9.539  < 2e-16 ***
as.factor(newzip)11224 -2.374e-01  6.343e-02  -3.743 0.000183 ***
as.factor(newzip)11225  7.070e-01  7.102e-02   9.955  < 2e-16 ***
as.factor(newzip)11231  9.289e-01  8.461e-02  10.979  < 2e-16 ***
as.factor(newzip)11235  2.932e-01  5.721e-02   5.125 3.02e-07 ***
as.factor(newzip)11238  9.102e-01  2.062e-01   4.414 1.02e-05 ***
as.factor(newzip)11249  1.996e+00  1.201e-01  16.618  < 2e-16 ***
as.factor(newzip)11354 -3.110e-01  1.411e-01  -2.204 0.027542 *
as.factor(newzip)11355 -1.654e-01  5.084e-02  -3.252 0.001148 **
as.factor(newzip)11357  4.692e-01  6.378e-02   7.357 2.00e-13 ***
as.factor(newzip)11358  2.911e-01  4.811e-02   6.050 1.49e-09 ***
as.factor(newzip)11364  2.287e-01  3.471e-02   6.590 4.60e-11 ***
as.factor(newzip)11366  1.866e-01  4.609e-02   4.049 5.17e-05 ***
as.factor(newzip)11367  2.785e-01  3.974e-02   7.009 2.54e-12 ***
as.factor(newzip)11368 -9.313e-01  1.080e-01  -8.627  < 2e-16 ***
as.factor(newzip)11369 -1.249e-01  4.859e-02  -2.571 0.010158 *
as.factor(newzip)11370  1.872e-01  6.075e-02   3.081 0.002069 **
as.factor(newzip)11373 -5.828e-01  7.242e-02  -8.048 9.27e-16 ***
as.factor(newzip)11374  3.964e-01  5.682e-02   6.977 3.19e-12 ***
as.factor(newzip)11375  4.343e-01  3.494e-02  12.430  < 2e-16 ***
as.factor(newzip)11379  3.778e-01  6.910e-02   5.468 4.64e-08 ***
as.factor(newzip)11385  4.181e-01  5.545e-02   7.540 5.04e-14 ***
as.factor(newzip)11413  8.497e-02  3.469e-02   2.450 0.014318 *
as.factor(newzip)11414 -1.684e-01  3.465e-02  -4.860 1.19e-06 ***
as.factor(newzip)11415  2.909e-01  9.459e-02   3.076 0.002105 **
as.factor(newzip)11433 -2.418e-01  3.500e-02  -6.908 5.17e-12 ***
log(LotSizeAcres)      1.303e-01  6.899e-03  18.887  < 2e-16 ***
log(LivingArea)        1.906e-01  8.241e-03  23.130  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4589 on 11536 degrees of freedom
Multiple R-squared:  0.5057,	Adjusted R-squared:  0.5017 
F-statistic: 124.2 on 95 and 11536 DF,  p-value: < 2.2e-16
```

Figure 18: Screenshot of Stepwise Model

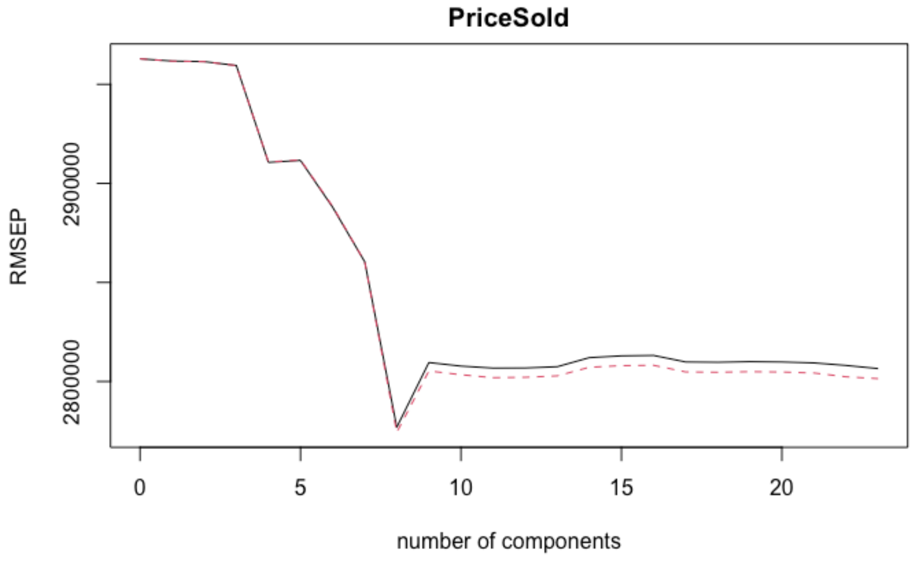


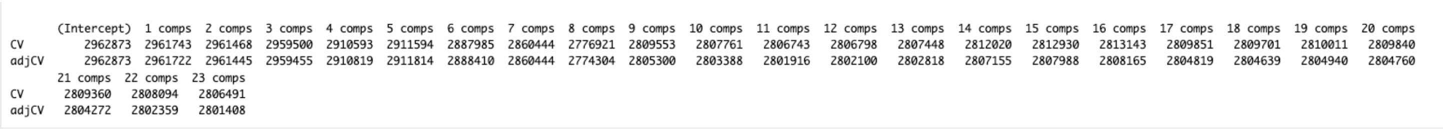
```
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
CV        2962873  2963016  2954753  2922622  2892930  2890922  2818921  2811776  2811545  2812186   2813707   2813662   2814614   2811821   2811690   2811869
adjCV     2962873  2962998  2954835  2922845  2893313  2891480  2817145  2808389  2807940  2808546   2809972   2809916   2810797   2807942   2807789   2807958
```



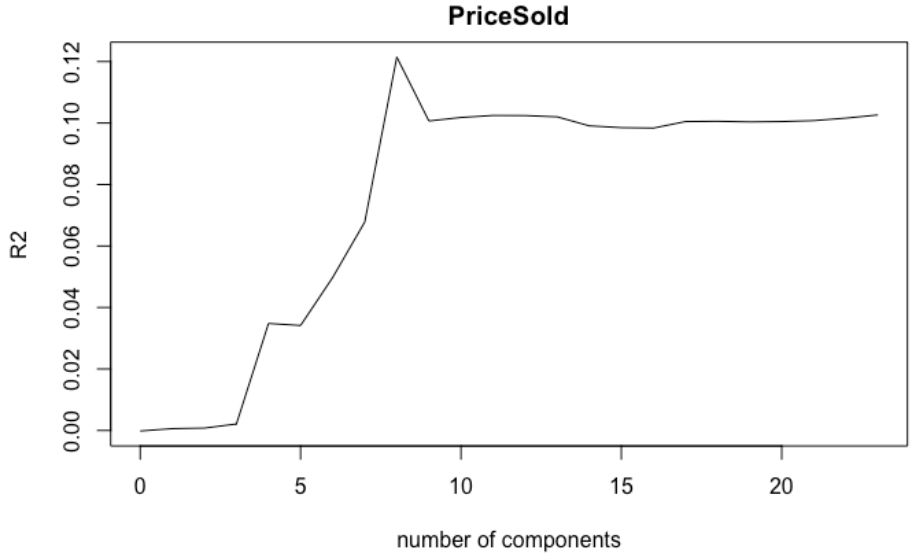
```
(Intercept)     1 comps     2 comps     3 comps     4 comps     5 comps     6 comps     7 comps     8 comps     9 comps    10 comps    11 comps    12 comps    13 comps    14 comps    15 comps
 -0.0001720  -0.0002684   0.0053027   0.0268185   0.0464916   0.0478148   0.0946545   0.0992379   0.0993861   0.0989753   0.0980006   0.0980294   0.0974189   0.0992090   0.0992934   0.0991785
```

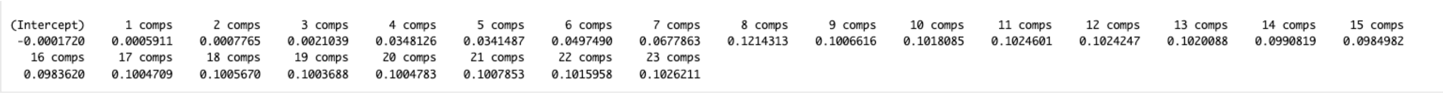
Figure 19: PCR Model 1





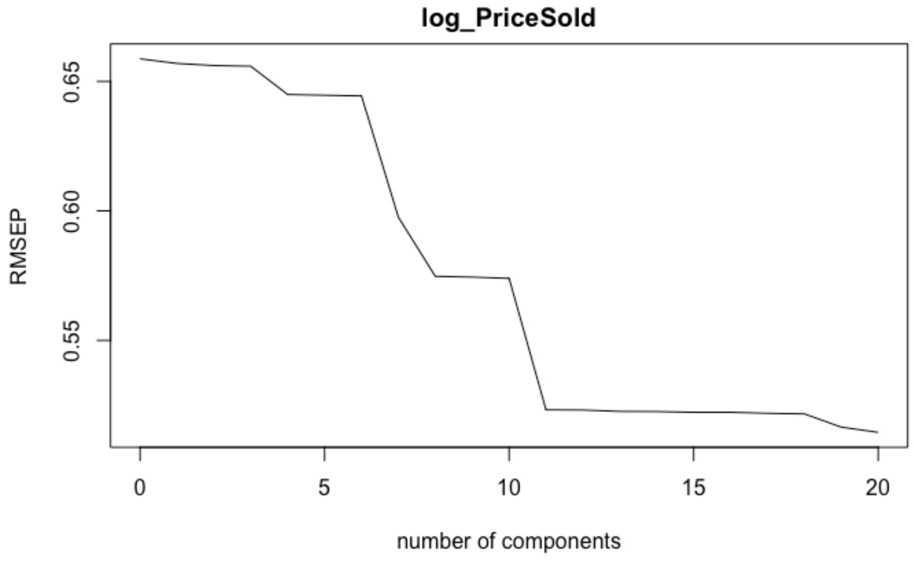
```
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
CV         2962873  2961743  2961468  2959500  2910593  2911594  2887985  2860444  2776921  2809553   2807761   2806743   2806798   2807448   2812020   2812930   2813143   2809851   2809701   2810011   2809840
adjCV      2962873  2961722  2961445  2959455  2910819  2911814  2888410  2860444  2774304  2805300   2803388   2801916   2802100   2802818   2807155   2807988   2808165   2804819   2804639   2804940   2804760
       21 comps  22 comps  23 comps
CV      2809360   2808094   2806491
adjCV   2804272   2802359   2801408
```



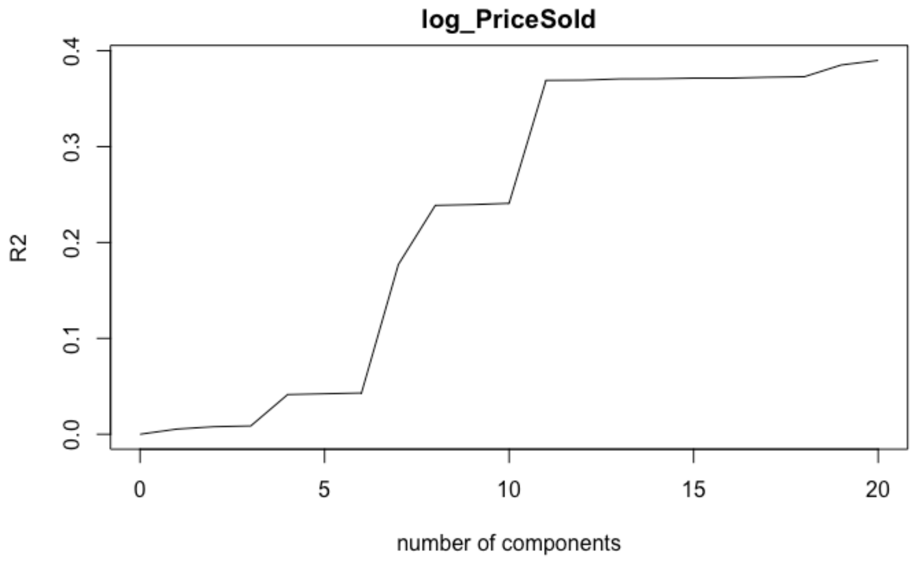


```
(Intercept)     1 comps     2 comps     3 comps     4 comps     5 comps     6 comps     7 comps     8 comps     9 comps    10 comps    11 comps    12 comps    13 comps    14 comps    15 comps 
 -0.0001720   0.0005911   0.0007765   0.0021039   0.0348126   0.0341487   0.0497490   0.0677863   0.1214313   0.1006616   0.1018085   0.1024601   0.1024247   0.1020088   0.0990819   0.0984982 
   16 comps    17 comps    18 comps    19 comps    20 comps    21 comps    22 comps    23 comps 
  0.0983620   0.1004709   0.1005670   0.1003688   0.1004783   0.1007853   0.1015958   0.1026211 
```

Figure 20: PCR Model 2

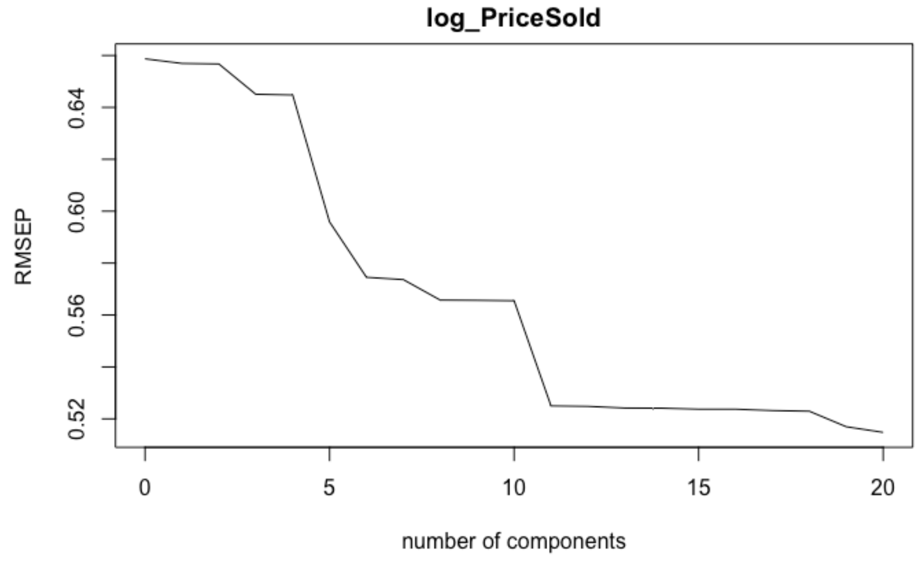


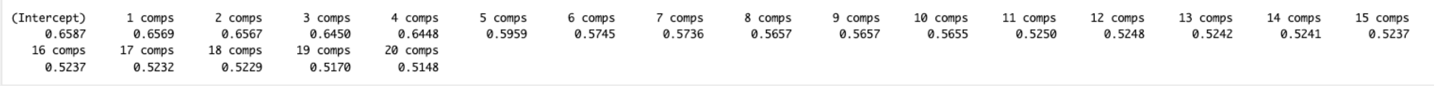
```
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps      6 comps      7 comps      8 comps      9 comps     10 comps     11 comps     12 comps     13 comps     14 comps     15 comps
     0.6587       0.6569       0.6561       0.6559       0.6449       0.6446       0.6444       0.5976       0.5747       0.5744       0.5739       0.5232       0.5231       0.5226       0.5225       0.5222
   16 comps     17 comps     18 comps     19 comps     20 comps
     0.5222       0.5219       0.5216       0.5165       0.5145
```



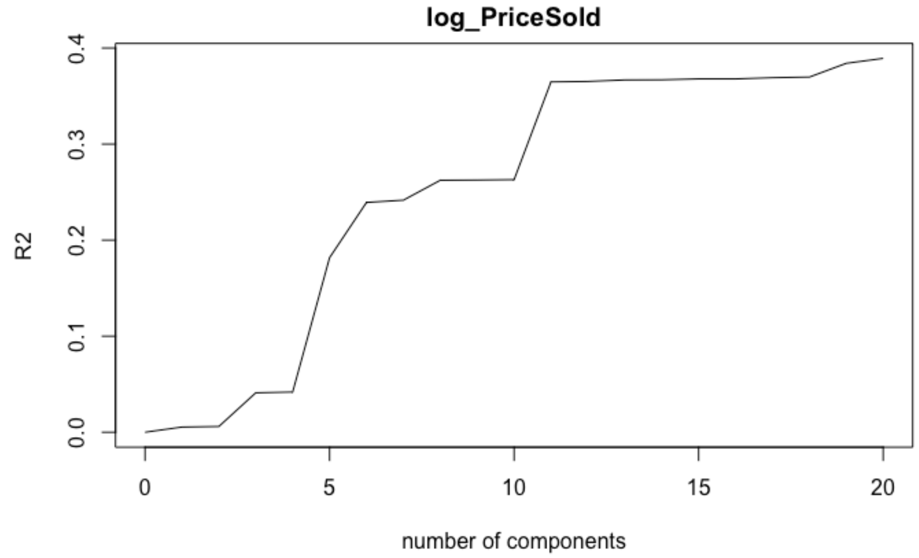
```
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps      6 comps      7 comps      8 comps      9 comps     10 comps     11 comps     12 comps     13 comps     14 comps     15 comps
   0.000000     0.005394     0.007862     0.008644     0.041385     0.042241     0.043022     0.177052     0.238680     0.239494     0.240847     0.369076     0.369303     0.370601     0.370737     0.371425
   16 comps     17 comps     18 comps     19 comps     20 comps
   0.371425     0.372353     0.372900     0.385118     0.389933
```

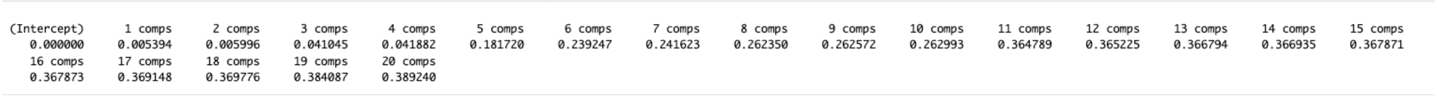
Figure 21: PCR Model 3





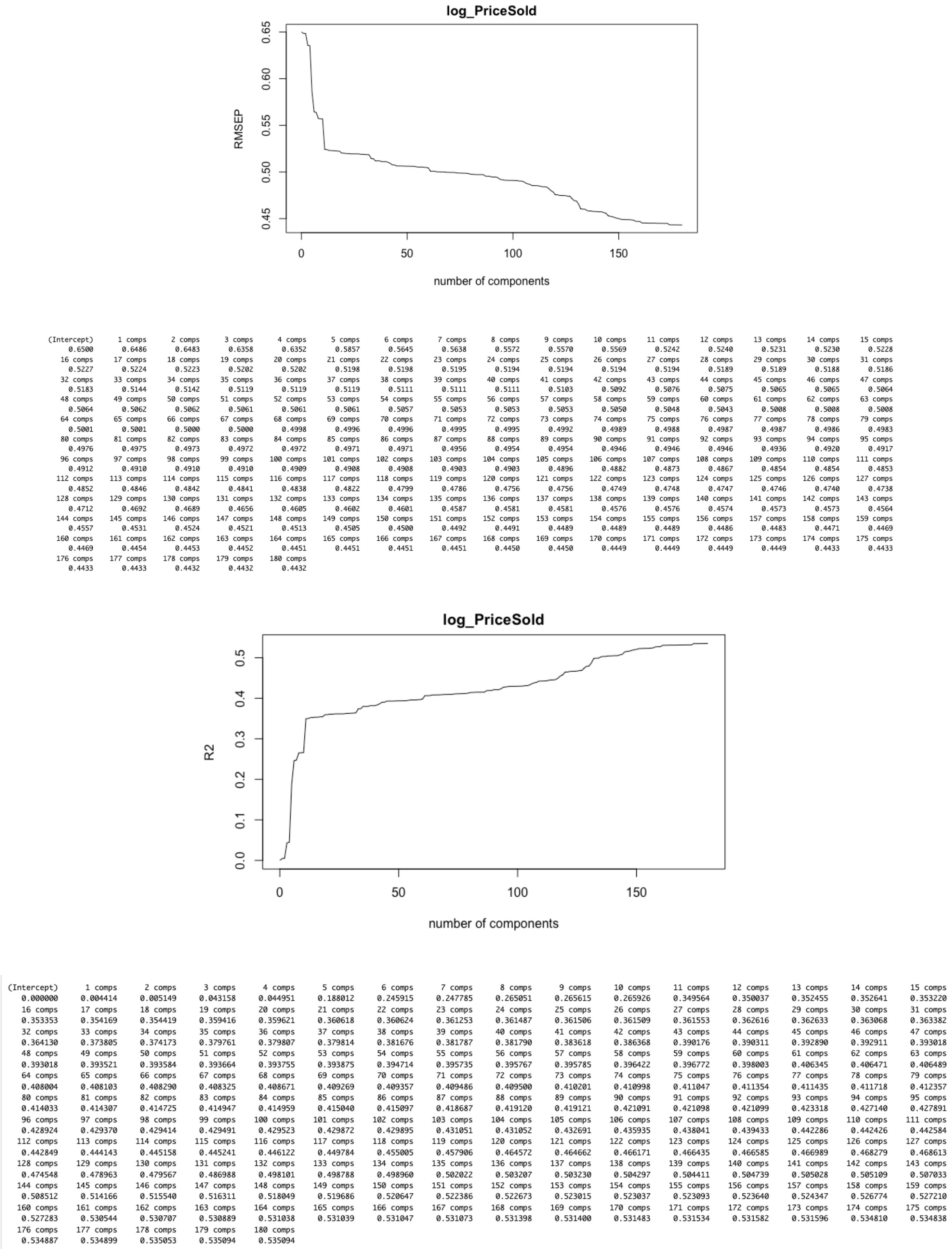
```
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps      6 comps      7 comps      8 comps      9 comps     10 comps     11 comps     12 comps     13 comps     14 comps     15 comps
     0.6587       0.6569       0.6567       0.6450       0.6448       0.5959       0.5745       0.5736       0.5657       0.5657       0.5655       0.5250       0.5248       0.5242       0.5241       0.5237
   16 comps     17 comps     18 comps     19 comps     20 comps
     0.5237       0.5232       0.5229       0.5170       0.5148
```





```
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps      6 comps      7 comps      8 comps      9 comps     10 comps     11 comps     12 comps     13 comps     14 comps     15 comps
   0.000000     0.005394     0.005996     0.041045     0.041882     0.181720     0.239247     0.241623     0.262350     0.262572     0.262993     0.364789     0.365225     0.366794     0.366935     0.367871
   16 comps     17 comps     18 comps     19 comps     20 comps
   0.367873     0.369148     0.369776     0.384087     0.389240
```

Figure 22: PCR Model 4



```
(Intercept)     1 comps     2 comps     3 comps     4 comps     5 comps     6 comps     7 comps     8 comps     9 comps    10 comps    11 comps    12 comps    13 comps    14 comps    15 comps
     0.6500      0.6486      0.6483      0.6358      0.6352      0.5857      0.5645      0.5638      0.5572      0.5570      0.5569      0.5242      0.5240      0.5231      0.5230      0.5228
   16 comps    17 comps    18 comps    19 comps    20 comps    21 comps    22 comps    23 comps    24 comps    25 comps    26 comps    27 comps    28 comps    29 comps    30 comps    31 comps
     0.5227      0.5224      0.5223      0.5202      0.5202      0.5198      0.5198      0.5195      0.5194      0.5194      0.5194      0.5194      0.5189      0.5189      0.5188      0.5186
   32 comps    33 comps    34 comps    35 comps    36 comps    37 comps    38 comps    39 comps    40 comps    41 comps    42 comps    43 comps    44 comps    45 comps    46 comps    47 comps
     0.5183      0.5144      0.5142      0.5119      0.5119      0.5119      0.5111      0.5111      0.5111      0.5103      0.5092      0.5076      0.5075      0.5065      0.5065      0.5064
   48 comps    49 comps    50 comps    51 comps    52 comps    53 comps    54 comps    55 comps    56 comps    57 comps    58 comps    59 comps    60 comps    61 comps    62 comps    63 comps
     0.5064      0.5062      0.5062      0.5061      0.5061      0.5061      0.5057      0.5053      0.5053      0.5053      0.5050      0.5048      0.5043      0.5008      0.5008      0.5008
   64 comps    65 comps    66 comps    67 comps    68 comps    69 comps    70 comps    71 comps    72 comps    73 comps    74 comps    75 comps    76 comps    77 comps    78 comps    79 comps
     0.5001      0.5001      0.5000      0.5000      0.4998      0.4996      0.4996      0.4995      0.4995      0.4992      0.4989      0.4988      0.4987      0.4987      0.4986      0.4983
   80 comps    81 comps    82 comps    83 comps    84 comps    85 comps    86 comps    87 comps    88 comps    89 comps    90 comps    91 comps    92 comps    93 comps    94 comps    95 comps
     0.4976      0.4975      0.4973      0.4972      0.4972      0.4971      0.4971      0.4956      0.4954      0.4954      0.4946      0.4946      0.4946      0.4936      0.4920      0.4917
   96 comps    97 comps    98 comps    99 comps   100 comps   101 comps   102 comps   103 comps   104 comps   105 comps   106 comps   107 comps   108 comps   109 comps   110 comps   111 comps
     0.4912      0.4910      0.4910      0.4910      0.4909      0.4908      0.4908      0.4903      0.4903      0.4896      0.4882      0.4873      0.4867      0.4854      0.4854      0.4853
  112 comps   113 comps   114 comps   115 comps   116 comps   117 comps   118 comps   119 comps   120 comps   121 comps   122 comps   123 comps   124 comps   125 comps   126 comps   127 comps
     0.4852      0.4846      0.4842      0.4841      0.4838      0.4822      0.4799      0.4786      0.4756      0.4756      0.4749      0.4748      0.4747      0.4746      0.4740      0.4738
  128 comps   129 comps   130 comps   131 comps   132 comps   133 comps   134 comps   135 comps   136 comps   137 comps   138 comps   139 comps   140 comps   141 comps   142 comps   143 comps
     0.4712      0.4692      0.4689      0.4656      0.4605      0.4602      0.4601      0.4587      0.4581      0.4581      0.4576      0.4576      0.4574      0.4573      0.4573      0.4564
  144 comps   145 comps   146 comps   147 comps   148 comps   149 comps   150 comps   151 comps   152 comps   153 comps   154 comps   155 comps   156 comps   157 comps   158 comps   159 comps
     0.4557      0.4531      0.4524      0.4521      0.4513      0.4505      0.4500      0.4492      0.4491      0.4489      0.4489      0.4489      0.4486      0.4483      0.4471      0.4469
  160 comps   161 comps   162 comps   163 comps   164 comps   165 comps   166 comps   167 comps   168 comps   169 comps   170 comps   171 comps   172 comps   173 comps   174 comps   175 comps
     0.4469      0.4454      0.4453      0.4452      0.4451      0.4451      0.4451      0.4451      0.4450      0.4450      0.4449      0.4449      0.4449      0.4449      0.4433      0.4433
  176 comps   177 comps   178 comps   179 comps   180 comps
     0.4433      0.4433      0.4432      0.4432      0.4432
```

```
(Intercept)     1 comps     2 comps     3 comps     4 comps     5 comps     6 comps     7 comps     8 comps     9 comps    10 comps    11 comps    12 comps    13 comps    14 comps    15 comps
   0.000000    0.004414    0.005149    0.043158    0.044951    0.188012    0.245915    0.247785    0.265051    0.265615    0.265926    0.349564    0.350037    0.352455    0.352641    0.353220
   16 comps    17 comps    18 comps    19 comps    20 comps    21 comps    22 comps    23 comps    24 comps    25 comps    26 comps    27 comps    28 comps    29 comps    30 comps    31 comps
   0.353353    0.354169    0.354419    0.359416    0.359621    0.360618    0.360624    0.361253    0.361487    0.361506    0.361509    0.361553    0.362616    0.362633    0.363068    0.363382
   32 comps    33 comps    34 comps    35 comps    36 comps    37 comps    38 comps    39 comps    40 comps    41 comps    42 comps    43 comps    44 comps    45 comps    46 comps    47 comps
   0.364130    0.373805    0.374173    0.379761    0.379807    0.379814    0.381676    0.381787    0.381790    0.383618    0.386368    0.390176    0.390311    0.392890    0.392911    0.393018
   48 comps    49 comps    50 comps    51 comps    52 comps    53 comps    54 comps    55 comps    56 comps    57 comps    58 comps    59 comps    60 comps    61 comps    62 comps    63 comps
   0.393018    0.393521    0.393584    0.393664    0.393755    0.393875    0.394714    0.395735    0.395767    0.395785    0.396422    0.396772    0.398003    0.406345    0.406471    0.406489
   64 comps    65 comps    66 comps    67 comps    68 comps    69 comps    70 comps    71 comps    72 comps    73 comps    74 comps    75 comps    76 comps    77 comps    78 comps    79 comps
   0.408004    0.408103    0.408290    0.408325    0.408671    0.409269    0.409357    0.409486    0.409500    0.410201    0.410998    0.411047    0.411354    0.411435    0.411718    0.412357
   80 comps    81 comps    82 comps    83 comps    84 comps    85 comps    86 comps    87 comps    88 comps    89 comps    90 comps    91 comps    92 comps    93 comps    94 comps    95 comps
   0.414033    0.414307    0.414725    0.414947    0.414959    0.415040    0.415097    0.418687    0.419120    0.419121    0.421091    0.421098    0.421099    0.423318    0.427140    0.427891
   96 comps    97 comps    98 comps    99 comps   100 comps   101 comps   102 comps   103 comps   104 comps   105 comps   106 comps   107 comps   108 comps   109 comps   110 comps   111 comps
   0.428924    0.429370    0.429414    0.429491    0.429523    0.429872    0.429895    0.431051    0.431052    0.432691    0.435935    0.438041    0.439433    0.442286    0.442426    0.442584
  112 comps   113 comps   114 comps   115 comps   116 comps   117 comps   118 comps   119 comps   120 comps   121 comps   122 comps   123 comps   124 comps   125 comps   126 comps   127 comps
   0.442849    0.444143    0.445158    0.445241    0.446122    0.449784    0.455005    0.457906    0.464572    0.464662    0.466171    0.466435    0.466585    0.466989    0.468279    0.468613
  128 comps   129 comps   130 comps   131 comps   132 comps   133 comps   134 comps   135 comps   136 comps   137 comps   138 comps   139 comps   140 comps   141 comps   142 comps   143 comps
   0.474548    0.478963    0.479567    0.486988    0.498101    0.498788    0.498960    0.502022    0.503207    0.503230    0.504297    0.504411    0.504739    0.505028    0.505109    0.507033
  144 comps   145 comps   146 comps   147 comps   148 comps   149 comps   150 comps   151 comps   152 comps   153 comps   154 comps   155 comps   156 comps   157 comps   158 comps   159 comps
   0.508512    0.514166    0.515540    0.516311    0.518049    0.519686    0.520647    0.522386    0.522673    0.523015    0.523037    0.523093    0.523640    0.524347    0.526774    0.527210
  160 comps   161 comps   162 comps   163 comps   164 comps   165 comps   166 comps   167 comps   168 comps   169 comps   170 comps   171 comps   172 comps   173 comps   174 comps   175 comps
   0.527283    0.530544    0.530707    0.530889    0.531038    0.531039    0.531047    0.531073    0.531398    0.531400    0.531483    0.531534    0.531582    0.531596    0.534810    0.534838
  176 comps   177 comps   178 comps   179 comps   180 comps
   0.534887    0.534899    0.535053    0.535094    0.535094
```

Figure 23: PCR Model 5