NH Real Estate Assessed Value Report

Phoebe Cheung, Margaret Luo, Pei Tao

October 1, 2017

1. Introduction

In this paper, we aim to build a model to predict the total value (`totval`) of properties in New Haven, Connecticut. The data used for this analysis was scraped from the Vision Government Solutions website, a database containing assessment information on properties in New Haven.

2. Methodology for Data Cleaning

Before any modeling, we cleaned our data to capture the most accurate longitudinal and latitudinal coordinates. Plotting the coordinates of each parcel ID (`pid`) and distinguishing the shape of New Haven, we were able to identify clusters of `pid`s clearly of the New Haven area. We used the `ggmaps` package to collect the longitudes and latitudes using the street addresses of these `pid`s. Once we replotted these corrected `pid`s, we reevaluated our plotted map of New Haven. There were a few more individual outliers to the map, whose longitudes and latitudes we reported as `NA`s.

3. Exploratory Data Analysis

3.1 Intuition

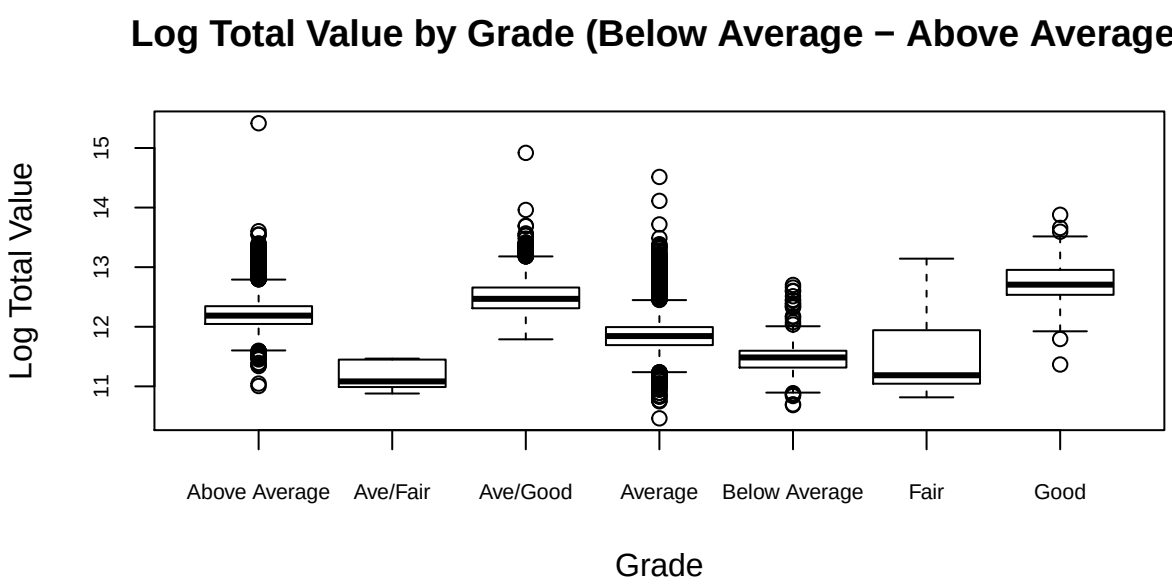
In order to gain a preliminary sense of which of the variables we scraped would be good predictors for `totval`, we eliminated a few that we wanted to look at. We first determined that owner address and previous sale history would not be helpful in determining the value of a property. Looking at the summaries for the variables helped us determine that bathroom style and kitchen style would not be good predictors because the majority of the `pid`s were evaluated as "Average" and less than three hundred total as either "Below Average" or "Above Average". We then determined from the summaries to not include garage square feet (`garagesqft`) because approximately half of the `pid`s contained `NA`s; we did not want to assume that `NA` implied no garage (a garage of 0 square footage). `Occupancy` and `model` both described how many occupants and families a property could house respectively, which we believed could be related to the number of bedrooms and bathrooms. We decided that the number of bedrooms, bathrooms, and half bathrooms (`halfbaths`) could be good predictors of the `totval`, and we used these variables instead.

3.2 Grouping *Grade*

Continuing to look at the summaries of the variables, we noticed that percent good (`pctgood`) and `grade` both measured the quality of the property. We decided to do some exploratory analysis on the categorical variable `grade`. Examining the summary for `grade`, we noticed that some categories had low counts. We chose to regroup the 16 categories into five more general ones, creating a new variable `grade2`. We consolidated the following:

- **4. Excellent**: "Very Good", "Vg/Exc", "Excellent", and "Excellent +"
- **5. Superior**: "Superior -", "Superior", "Superior +", and "Luxurious"

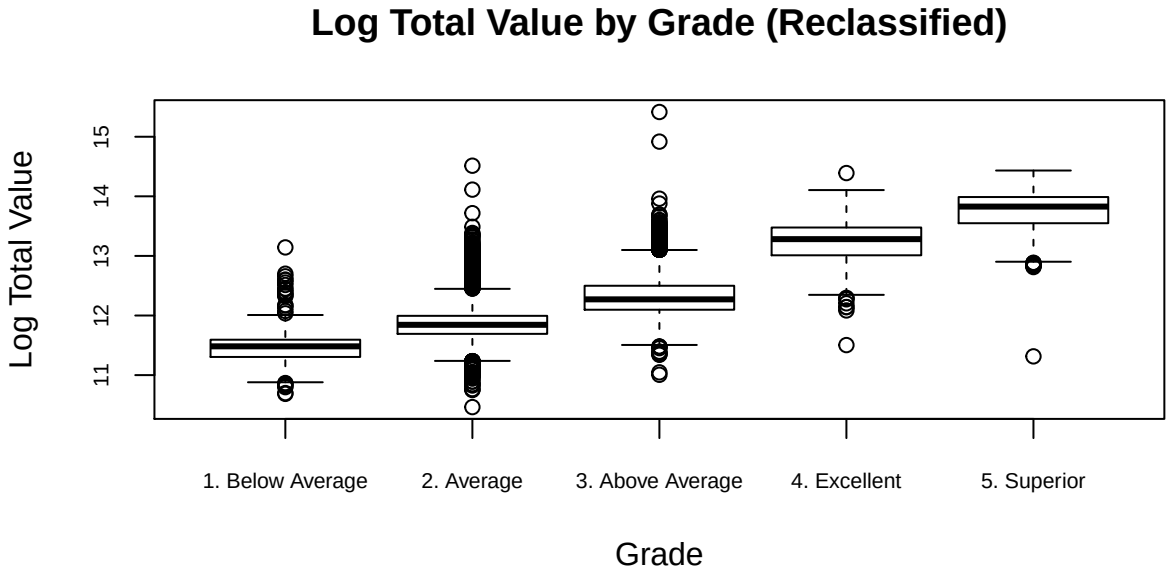
In order to understand the differences between "Average", "Fair", "Good", "Below Average", and the various combinations of these classifications, we looked at a boxplot of these categories and based our groupings on their distributions of `totval`:



Using this boxplot as guidance we made the following reclassifications:

- **1. Below Average**: "Low Cost", "Below Average", "Fair" and "Ave/Fair"
- **2. Average**: "Average"
- **3. Above Average**: "Good", "Ave/Good", and "Above Average"

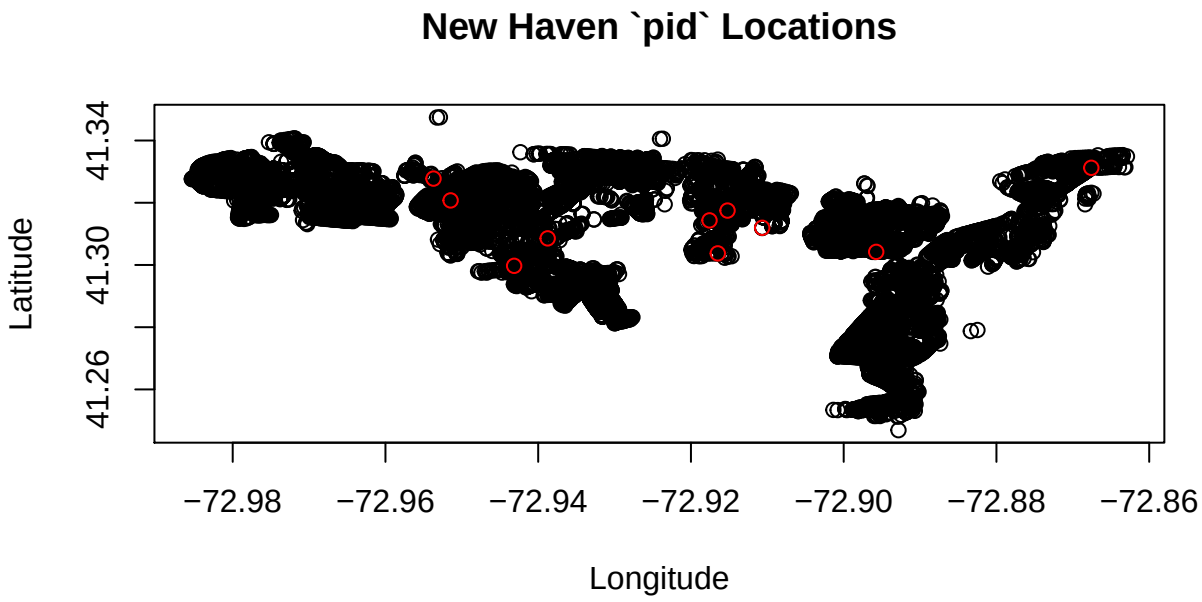
We decided to keep "Average" as it was without grouping it with anything else, since its boxplot showed many outliers. We decided this could be due to different evaluators' interpretation of the generic term "average". By using this new grouping, we can see the distributions of `totval` by classification:



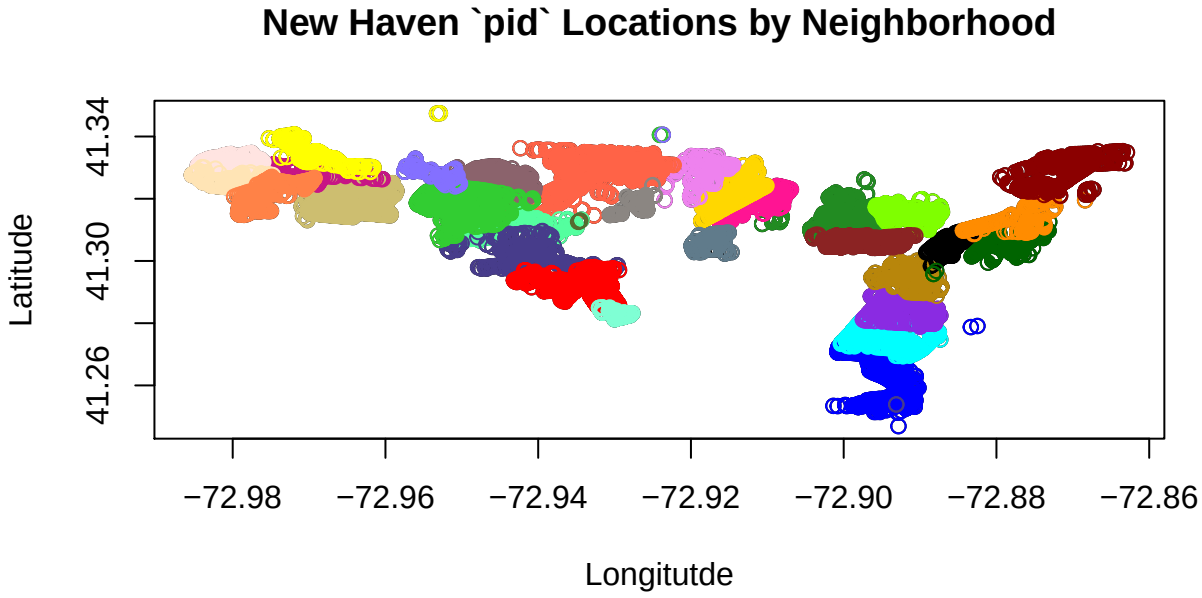
We can clearly see that the distribution of `totval` centers among higher values of `totval` as the `pid`s increase in grade evaluation, which may mean that `grade` would be a good predictor of `totval`.

3.3 Grouping `Neighborhood`

Once again looking at the summary of the variables, we noticed neighborhoods to have low counts the categories "CHP3", "CHP4", "G1", "IND5", "OR3", "Q1", "W", "WHA4", and "X", which accounted for 10 `pid`s in total. When plotting the `pid`s in these neighborhoods on our latitude vs. longitude plot, we noticed that they were quite spread out:



We decided to not include these categories for neighborhoods and changed them to `NA`'s in a new variable `neighborhood2`. We then reproduced our latitude vs. longitude plot by color coding by neighborhood:



By seeing that each `pid` seems clustered by a neighborhood in New Haven, we believed that `neighborhood2` would be a good predictor of `totval` and would implicitly include longitudinal and latitudinal data in the analysis.

3.4 Other Variables Considered

Other variables besides `bedrooms`, `bathrooms`, `halfbaths`, our reclassified `grade2`, and our reclassified `neighborhood2` that we wanted to consider for our analysis were `yearbuilt`, `sqft`, `acres`, `style`, and `actype`. We believed that when a property was built might affect its value, as we felt that properties tend to grow in value over time. However, we also considered the case that an older, improperly-maintained property might lose value as it falls into disrepair. If a property had more livable square feet and acres of land, we believed that it would have higher value. In addition, we hypothesized that the existence and type of air conditioning (`actype`) would affect the overall property value. In the case of `style`, we considered the idea that different housing styles may be more attractive and easier to properly maintain, increasing the property's overall value.

4. Regression Analysis

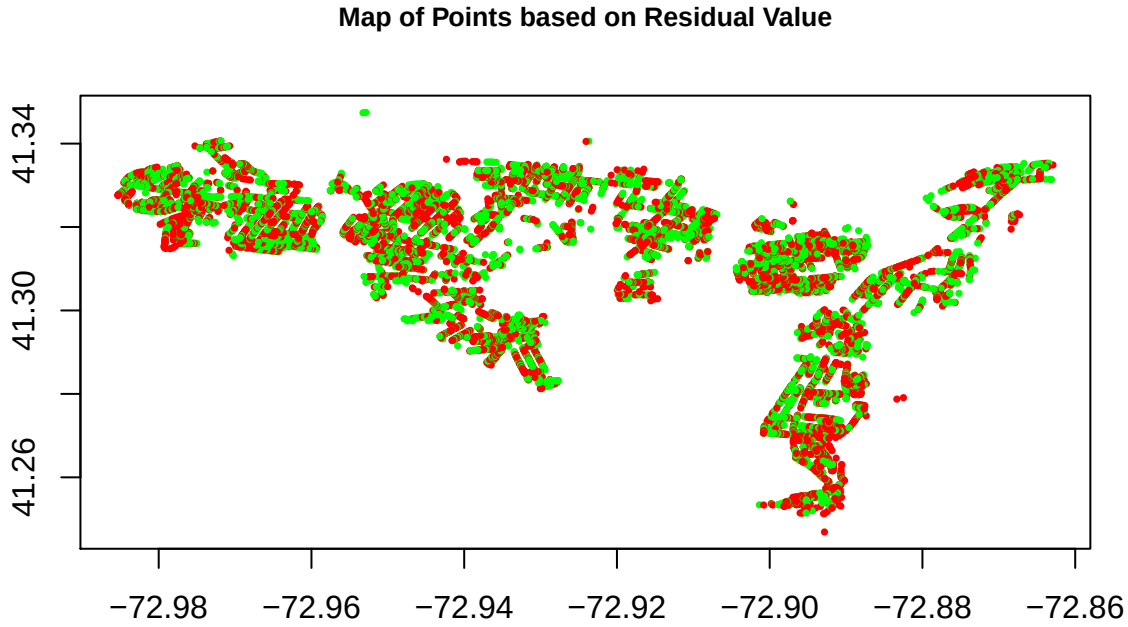
4.1 The Model

We decided to run a linear model with different combinations of the variables we determined beforehand to possibly estimate `totval`. We chose to take the logarithm of `totval` due to the magnitude of the values and the fact that the distribution is right skewed. The model that we found to best estimate `totval` is as follows:

```
MFinal <- lm(log(totval) ~ bedrooms + bathrooms + halfbaths + acres + sqft +
               neighborhood2 + grade2 + actype, data = MasterData)
summary(MFinal)$r.squared
```

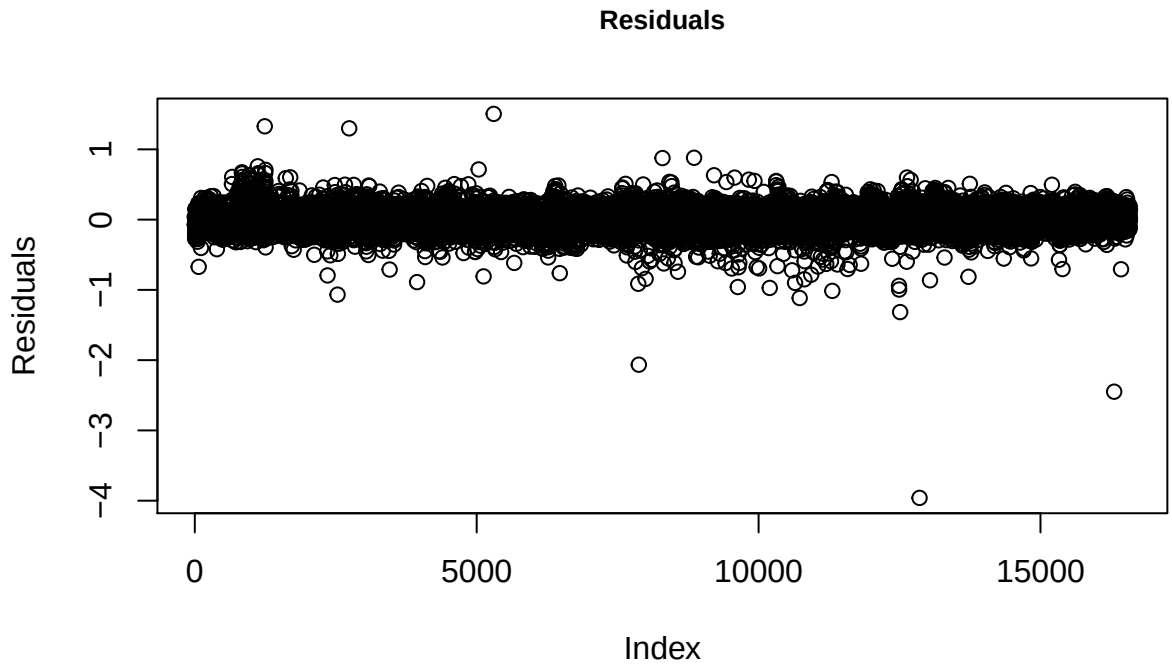
```
## [1] 0.903944
```

See Appendix A for the full summary of the linear model. After running the regression, we visualized the residuals in two different ways:



In this plot of the residuals by longitudinal and latitudinal locations, the green points represent `pid`'s with

positive residuals, red represent negative. Had there been clustering of red and green, this would mean some neighborhoods were poorly represented by the model. However, since we do not see this effect, the result is that we believe our model represents the neighborhoods well.



Our residuals vs. fitted plot looks relatively well-behaved with no obvious trends and only a few outliers, suggesting that our choice of model to predict `totval` was appropriate.

4.2 Model Interpretation

After testing various linear models, we found that `yearbuilt` was not a statistically significantpredictor of `totval` (at the $\alpha = 0.05$ level). This may be due to our beliefs mentioned above that an older building could have increased or decreased over time depending on how its maintainence over the years.

Our model implies that, as one might expect, the number of bedrooms and bathrooms and the acreage and square footage of a house were positively associated with its value, i.e., when acreage or the number of bedrooms increases, so does the house's value. Certain neighborhoods (e.g. neighborhoods 21300, 21400) were also positively associated with a house's value, whereas others (e.g. neighborhoods 20200, 20300) were negatively associated.

Almost all of the variables we included in our model had highly significant coefficients, suggesting that we had intuited correctly in our variable choice. Our residual plots also show a promising "patternlessness," so to speak. Based on the initial analysis, the testing of various regressions with combinations of variables to optimize the predictability, and the resulting seemingly well-behaved residuals, we found that the linear model containing the combination of `bedrooms`, `bathrooms`, `halfbaths`, `acres`, `sqft`, `neighborhood2`, `grade2`, and `actype` best predicted the total value of a property in New Haven, Connecticut.

Appendix A

```
summary(MFinal)
```

```
##
## Call:
## lm(formula = log(totval) ~ bedrooms + bathrooms + halfbaths +
##     acres + sqft + neighborhood2 + grade2 + actype, data = MasterData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9602 -0.0853  0.0006  0.0862  1.5047
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.167e+01  3.268e-02 357.050  < 2e-16 ***
## bedrooms                 1.652e-02  1.149e-03  14.379  < 2e-16 ***
## bathrooms                6.420e-02  1.880e-03  34.154  < 2e-16 ***
## halfbaths                7.902e-02  2.744e-03  28.800  < 2e-16 ***
## acres                    1.521e-01  1.206e-02  12.608  < 2e-16 ***
## sqft                     8.644e-05  1.555e-06  55.579  < 2e-16 ***
## neighborhood20200       -4.887e-02  8.099e-03  -6.034 1.64e-09 ***
## neighborhood20300       -1.834e-01  8.089e-03 -22.677  < 2e-16 ***
## neighborhood20400       -2.412e-01  8.995e-03 -26.810  < 2e-16 ***
## neighborhood20500       -2.883e-01  9.795e-03 -29.430  < 2e-16 ***
## neighborhood20600       -2.174e-01  1.065e-02 -20.421  < 2e-16 ***
## neighborhood20700       -2.886e-01  1.020e-02 -28.290  < 2e-16 ***
## neighborhood20800       -1.930e-01  7.774e-03 -24.822  < 2e-16 ***
## neighborhood20900       -4.887e-01  7.596e-03 -64.332  < 2e-16 ***
## neighborhood21000       -2.473e-01  8.310e-03 -29.761  < 2e-16 ***
## neighborhood21100       -3.691e-01  8.408e-03 -43.903  < 2e-16 ***
## neighborhood21200        5.729e-01  7.746e-03  73.954  < 2e-16 ***
## neighborhood21300        6.085e-01  9.055e-03  67.200  < 2e-16 ***
## neighborhood21400        2.109e-01  1.188e-02  17.747  < 2e-16 ***
## neighborhood21500        5.853e-01  1.360e-02  43.037  < 2e-16 ***
## neighborhood21600       -4.075e-01  6.440e-03 -63.273  < 2e-16 ***
## neighborhood21650        1.130e-01  1.250e-02   9.045  < 2e-16 ***
## neighborhood21700       -1.097e-01  7.955e-03 -13.788  < 2e-16 ***
## neighborhood21800       -3.463e-01  8.966e-03 -38.618  < 2e-16 ***
## neighborhood21801        3.846e-01  5.379e-02   7.150 9.04e-13 ***
## neighborhood21900       -3.598e-01  7.897e-03 -45.559  < 2e-16 ***
## neighborhood22000       -4.970e-01  7.060e-03 -70.387  < 2e-16 ***
## neighborhood22100       -1.566e-01  1.504e-02 -10.414  < 2e-16 ***
## neighborhood22200       -1.901e-01  7.188e-03 -26.442  < 2e-16 ***
## neighborhood22300       -2.223e-01  1.460e-02 -15.227  < 2e-16 ***
## neighborhood22400        1.211e-01  7.000e-03  17.302  < 2e-16 ***
## neighborhood22500       -1.685e-01  9.542e-03 -17.654  < 2e-16 ***
## neighborhood22600       -2.085e-01  9.855e-03 -21.156  < 2e-16 ***
## neighborhood22700       -1.771e-01  9.548e-03 -18.543  < 2e-16 ***
## neighborhood22800       -1.399e-01  9.387e-03 -14.899  < 2e-16 ***
## neighborhood22900        4.870e-02  9.093e-03   5.356 8.63e-08 ***
## grade22. Average         2.768e-01  7.838e-03  35.321  < 2e-16 ***
## grade23. Above Average   4.353e-01  8.306e-03  52.403  < 2e-16 ***
## grade24. Excellent       7.038e-01  1.135e-02  62.035  < 2e-16 ***
```

```
## grade25. Superior       8.887e-01  1.872e-02  47.468  < 2e-16 ***
## actypeCentral          -1.379e-01  3.137e-02  -4.398 1.10e-05 ***
## actypeHeat Pump        -1.240e-01  5.005e-02  -2.477   0.0133 *
## actypeNone             -2.083e-01  3.121e-02  -6.673 2.59e-11 ***
## actypePartial          -1.544e-01  3.697e-02  -4.177 2.97e-05 ***
## actypeUnit/AC          -7.044e-02  5.005e-02  -1.407   0.1593
## actypeVapor Cooler     -1.755e-01  1.115e-01  -1.573   0.1157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1513 on 16541 degrees of freedom
##   (232 observations deleted due to missingness)
## Multiple R-squared:  0.9039, Adjusted R-squared:  0.9037
## F-statistic:  3459 on 45 and 16541 DF,  p-value: < 2.2e-16
```