<https://www.freeimages.com/search/apartment>

**Executive Summary**

This project aimed to predict whether a person was satisfied or dissatisfied with their housing using their responses to the American Housing Survey.

The vast majority of people tended to be ‘satisfied’ with their housing. Surprisingly, the factor correlated with housing satisfaction the most was whether or not a person was satisfied with their neighborhood.

Outside of neighborhood satisfaction, people tended to be more strongly affected by factors that can be considered “temporary,” such as whether or not there is peeling paint, rather than factors that are permanent. Nine out the fifteen top factors correlated to housing satisfaction are features of a home that can be repaired or changed. This indicates that property managers can actually affect housing satisfaction levels within their own buildings.

A logistic regression model was used for predicting a binary response variable, where 1 equaled satisfaction with housing and 0 equaled dissatisfaction. The greatest accuracy achieved was…? While a random forest model was also attempted, tuning the hyperparameters took considerably more time and computing power, with no increase in model accuracy.

**BACKGROUND**

**Purpose**

Housing preferences in the United States have changed notably over the last several decades for a multitude of reasons, such as increased density in urban cores, changes in the average family size, and shifting trends in architectural styles.

The intent of this study is to elucidate which factors affect housing satisfaction the most and examine whether or not there are variations in housing preferences from region to region. Possible factors influencing housing satisfaction include, but are not limited to: home size, cost of housing, quality of the home, or age of the home. These can all be considered “intrinsic” to each home after construction. Other “temporary” factors may also affect housing satisfaction, such as the presence of peeling paint, whether or not the home has a dishwasher, or if the home’s exterior is damaged.

Finally, this study will use a machine learning model to predict whether a survey respondent is satisfied or dissatisfied with their housing using the other survey questions.

**Intended Audience**

There are several groups who would be able to leverage this data to make informed decisions:

**Housing Developers:** This data could help developers create high level strategies for new housing developments, as well as aid in decision-making for purchasing existing housing developments.

**Large-Scale Property Owners:** Property owners could tailor their renovation schedules to reduce the risk of tenant churn over time, and prioritize changing housing units which have characteristics that make it likely for tenants to be dissatisfied.

**Municipalities:** City redevelopment agencies or neighborhood community development groups are inclined to incentivize housing developments that are appealing to residents, and as a result, could use this data to guide their policies and development initiatives to promote resident retention.

**Data Source**

The data used for this project is from the 2015 American Housing Survey (AHS). This data is available at the following URL: <https://catalog.data.gov/dataset/american-housing-survey-ahs>.

The AHS national survey was conducted annually from 1973-1981 and every two years from 1983 – 2015. The survey questions have changed over time, but the questions have always pertained to the characteristics of the housing unit, such as the size and composition, as well as characteristics of the tenants and the survey respondents, such as their age, race, or income level.

For the purposes of this project, only characteristics directly related to the housing unit and the surrounding neighborhood were used, and questions pertaining to the socioeconomic background of the respondent, their family, and their fellow tenants were disregarded.

The survey question used as the target variable in the logistic regression model is “Overall Opinion of Present Home (1 – 10).” The responses were split into two categories, where 1 through 5 are classified as 0, or dissatisfied, and 6 through 10 are classified as 1, or satisfied.

The original dataset contained many more fields than were needed, had quite a few missing values, and was stored as an encoded csv file. The data wrangling steps taken to clean the data up and make it usable for the machine learning model is contained here: ???

**EXPLORATORY DATA ANALYSIS**

**Visual Exploratory Data Analysis**

A few simple plots will be created using seaborn to examine and understand the data. The github repository containing the code used to generate all of the visualizations can be found here: ???

*Figure 1* is a bar plot depicting the distribution of home ratings. The most common rating that respondents gave their homes was a 10, followed by 8. Very few people gave their homes a rating lower than 5. As the model is trying to predict one of two categories, the uneven distribution of the target variable may make it tricky.

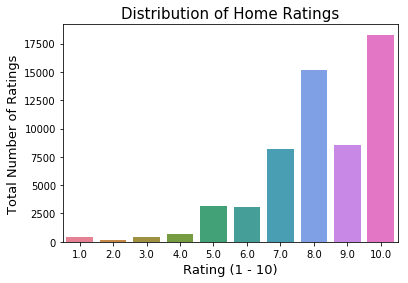


Figure 1

When home ratings are divided by metropolitan region, it is possible to see if there is a significant difference in housing satisfaction by region. The 2015 survey only included these sixteen regions and a category for “non-metropolitan” regions. The median rating for almost every region is 8, but the range of ratings varies slightly by region. The range of ratings for DC, Boston, and Atlanta were smallest, with 95% of ratings falling between 5 and 10. Seattle’s distribution of ratings is also slightly different. It may be worth examining the differences between ratings in metropolitan regions more closely.

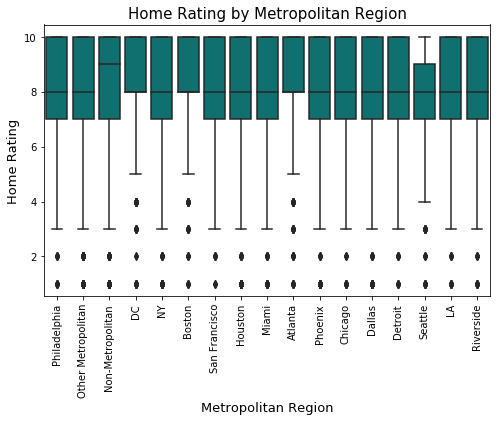
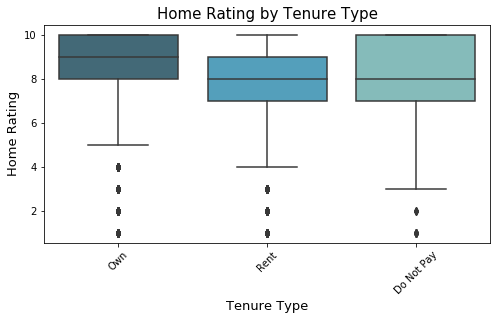


Figure 2

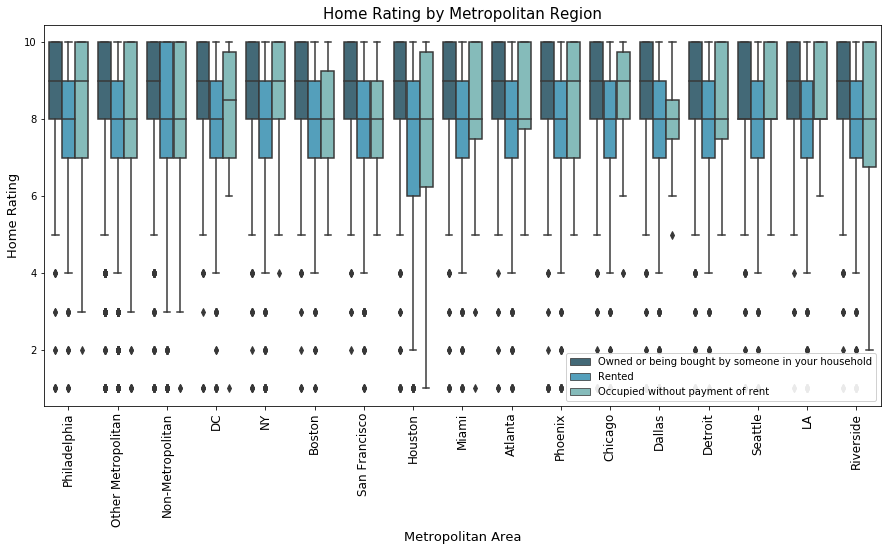
Do home ratings vary between renters and homeowners?  *Figure 3* shows home ratings divided by renters, owners, and survey respondents who do not pay for their housing, such as people who live in a friend’s home.

Homeowners tended to rate their homes higher than respondents who either rent their housing or do not pay for their housing, as shown in *Figure 3*.



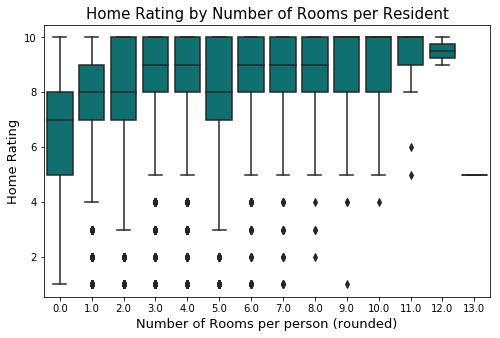
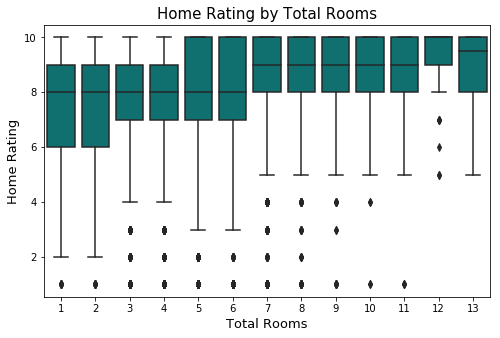
*Figure 3*

If we split the ratings for every tenure type into their respective metropolitan regions, it can be seen that renters and homeowners tend to rate their homes similarly regardless of metro region.



*Figure 4*

The number of rooms also may affect home ratings. *Figures 5 and 6* depict home ratings by the number of rooms in the home and the number of rooms per resident. Homes with more than six rooms tend to have the highest average and median ratings, as did homes with more than two rooms per resident.



*Figure 5 and Figure 6*

Finally, *Figure 7* shows home rating by when the home was built. There was a distinct gap in ratings in homes constructed prior to 1990.

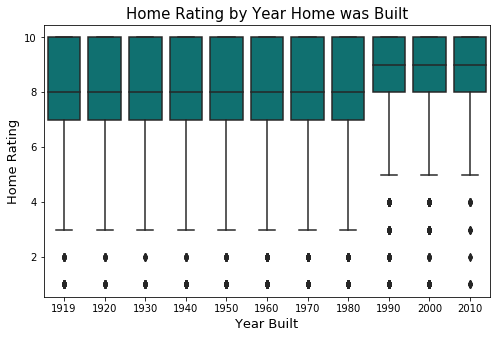


Figure 7

The neighborhood ratings were also charted onto a bar plot, and it shows a very similar distribution as the home ratings—10 is the most common rating, followed by 8, as depicted in *Figure 8*.

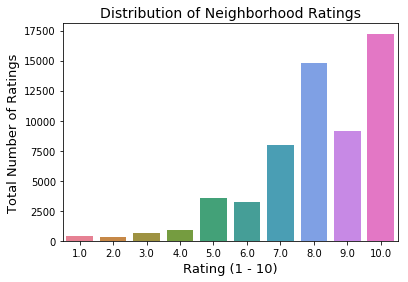


Figure 8

Neighborhood ratings are generally similar amongst the different metropolitan regions, but Boston and Non-Metropolitan neighborhoods had the highest median and average ratings.

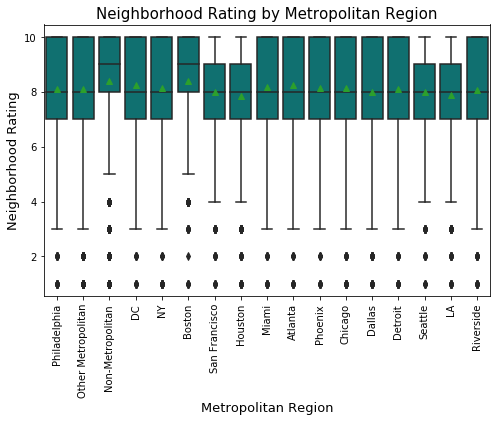


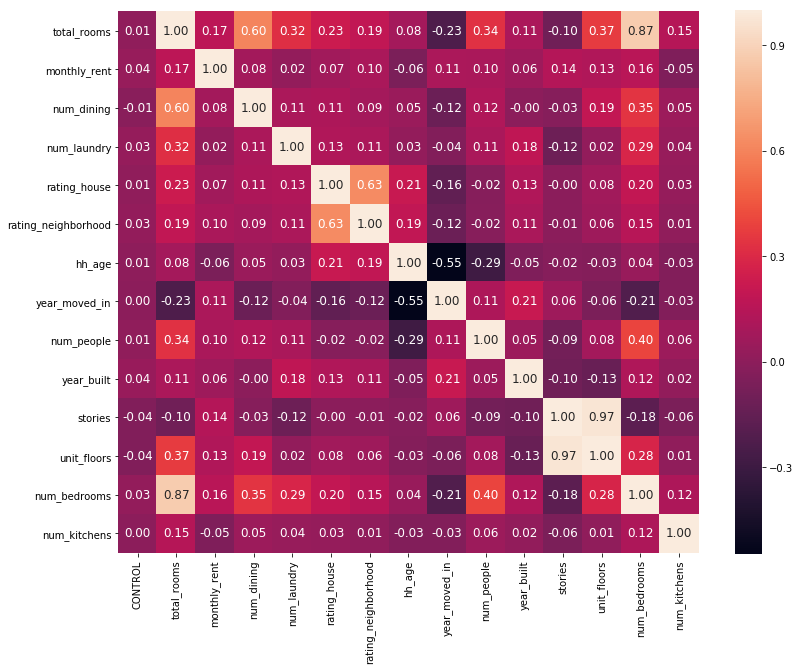
Figure 9

**Statistical Analysis**

Statistical methods will be applied to determine which features are most strongly related to the target variable. These methods will also guide feature selection if it is needed for the machine learning model.

Correlation Matrix

For the numerical features, a correlation matrix was computed. The majority of features had very low correlations with the target feature, rating\_house. Very few features had a correlation higher than 0.20. The entire correlation matrix is depicted below:



The values with the highest correlation with the target variable were:

* rating\_neighborhood (0.63015795903725946, p = 0.0)
* total\_rooms (0.23025241547195274, p = 0.0)
* hh\_age (0.2137715946399161, p = 0.0)
* num\_bedrooms (0.19557474849364975, p = 0.0)

All of the other correlations were under 0.20, but still have p-values under 0.05 (most of the p-values are 0). This indicates that the correlations are all statistically significant. However, the low Pearson correlations may make it difficult to use any of the numerical values in the machine learning model.

For the categorical variables, an analysis of variance (ANOVA) was performed to determine if any of the categorical variables contain a statistically significant difference between groups when looking at the home ratings.

The entire table of f-values and their p-values is attached in Appendix B, but the top 15 values are as follows:

|  |  |  |
| --- | --- | --- |
| **Feature** | **f-value** | **p-value** |
| petty\_crime | 2,642.07 | 0.00E+00 |
| serious\_crime | 2,449.85 | 0.00E+00 |
| rating\_neighborhood | 2,106.82 | 0.00E+00 |
| dishwash | 1,746.21 | 0.00E+00 |
| wall\_crack | 1,611.99 | 0.00E+00 |
| garage | 1,492.11 | 2.470328e-322 |
| paint\_peeling | 1,332.31 | 1.98E-288 |
| good\_schools | 1,309.70 | 2.49E-283 |
| washer | 1,168.80 | 1.21E-253 |
| near\_trash | 1,107.91 | 0.00E+00 |
| number\_upkeep\_probs | 1,056.57 | 0.00E+00 |
| adequacy | 1,028.01 | 0.00E+00 |
| floor\_hole | 986.54 | 9.52E-215 |
| tenure | 875.32 | 0.00E+00 |
| home\_better\_than\_last | 804.17 | 0.00E+00 |

|  |  |  |
| --- | --- | --- |
| Feature | X2 | p-value |
| num\_bathrooms | 3855.34 | 0.00E+00 |
| rent\_subsidy | 3295.41 | 0.00E+00 |
| adequacy | 2571.49 | 0.00E+00 |
| roach | 2409.25 | 0.00E+00 |
| unit\_size | 2078.07 | 0.00E+00 |
| wall\_crack | 1470.20 | 0.00E+00 |
| hud\_subsidized | 1385.21 | 3.43E-303 |
| roof\_sag | 1326.01 | 2.52E-290 |
| rodent | 1305.28 | 8.06E-286 |
| roof\_hole | 1300.68 | 8.02E-285 |
| missing\_siding | 1276.19 | 1.69E-279 |
| paint\_peeling | 1272.18 | 1.26E-278 |
| missing\_shingle | 1250.17 | 7.62E-274 |

**MACHINE LEARNING**

To predict whether or not a survey respondent is satisfied with their housing, a machine learning model was used. First, a binary response variable was created, where home ratings of 1 to 5 are classified as 0 and ratings of 6 to 10 are classified as 1. Out of the 58,233 samples, 53,285 are classified as 1, and 4,950 are classified as 0.

The features and the response variable are split into training and test sets, with the size of the training set as 70% of the samples.

**Naïve Model**

As the majority of respondents rated their homes above 5, a dummy model predicting the most frequent class was created. This model attained a fairly high accuracy overall, but has a recall of 0 for predicting the negative class.

[[ 0 1449]

[ 0 16021]]

precision recall f1-score support

0 0.00 0.00 0.00 1449

1 0.92 1.00 0.96 16021

avg / total 0.84 0.92 0.88 17470

Accuracy Score: 0.92

**Out of Box Logistic Regression**

First, a logistic regression model was trained with the default parameters. This model already has a slightly higher overall accuracy than the naïve model, but the recall and precision for the negative class is low.

[[ 396 1053]

[ 227 15794]]

precision recall f1-score support

0 0.64 0.27 0.38 1449

1 0.94 0.99 0.96 16021

avg / total 0.91 0.93 0.91 17470

**Tuned Logistic Regression**

Tuning the hyperparameter C may increase the model’s performance. C values of 0.0001, 0.001, 0.01, 0.1, 1, 10, and 100 were tested using cross validation. A C value of 100 resulted in the highest accuracy. Because lower C values result in stronger regularization, it seems that less regularization results in better model performance in this case.

[[ 396 1053]

[ 227 15794]]

precision recall f1-score support

0 0.64 0.27 0.38 1449

1 0.94 0.99 0.96 16021

avg / total 0.91 0.93 0.91 17470

Although the precision and recall didn’t increase, the tuned model is more likely to perform better for new data than an untuned model.

**Tuned Logistic Regression with Feature Selection**

Feature selection was performed to examine whether removing features could improve model performance. Features were ranked by their X2 statistic and removed if they fell below a certain threshold.

However, the model’s performance actually declined when features were removed. The more features removed, the lower the model’s accuracy, although only slightly. If model complexity were a concern, features can be removed with a small loss in accuracy, but for this model, all of the features were left in the model.

[[ 388 1061]

[ 235 15786]]

precision recall f1-score support

0 0.62 0.27 0.37 1449

1 0.94 0.99 0.96 16021

avg / total 0.91 0.93 0.91 17470

**Tuned Logistic Regression with Undersampling**

Because the response variable has only two classes and the positive class is present with an 11 to 1 ratio, every model is much better at predicting the modal class, 1, than it is at predicting the negative class. If the goal of the model is to perform equally as well for both classes, the dataset should be undersampled so that both response variables are present in equal amounts.

The dataset was undersampled so that both classes had 9,900 records, and then split into training and test sets. The hyperparameter C was tuned as well, and had the best results with a C value of 0.01.

The overall accuracy for the model decreased to 0.80, but the model performed equally well in predicting the positive class as it did in predicting the negative class.

[[1199 311]

[ 270 1190]]

precision recall f1-score support

0 0.82 0.79 0.80 1510

1 0.79 0.82 0.80 1460

avg / total 0.80 0.80 0.80 2970

**Random Forest Model**

A couple random forest models were also trained to

**CONCLUSION**

A simple logistic regression model resulted in the highest success in predicting the binary response variable. A tuned logistic regression model resulted in

* What affects the model the most
* Difficulty in Predicting
* Study Improvements

**APPENDIX A**

Data Wrangling­

Overview

The data used for this project is from the 2015 American Housing Survey. As the survey has far more attributes than are needed for an initial examination, we pull out the 130 most relevant attributes and examine those. The chosen attributes are

**1.0 Initial Impression**

The dataset has almost 70,000 entries and 1,091 attributes for each. The column names are abbreviated and make it difficult to understand what each attribute is. In addition, the values in the DataFrame are all encoded with numbers stored as strings. We will have to pull out the relevant columns, map all of the original values back into the DataFrame, and rename the columns so that it is clear what each attribute is when the analysis is performed.

**2.0 Files**

For this project, we need the original csv with all of the survey answers as well as two other files which are used to subset the data and rename values:

household.csv: This file holds all of the survey answers to the 2015 American Housing Survey

Column Values.xlsx: This excel file has the 129 columns that will be used in the analysis, and the new names for the columns.

AHS 2015 Value Labels.csv: This csv has all of the values and what they should be mapped to.

**3.0 Data Wrangling**

The household.csv file is read in as df\_household. It has 69,493 entries and 1,091 columns. To select the 129 columns that will be used in the analysis, ‘Column Values.xlsx’ is read in as df\_colnames and then used to select the columns from df\_household. The new DataFrame has 129 columns and is named df\_household\_clean.

The .strip() method is used to remove all of the single quotation marks which surround all of the values in the DataFrame.

The ‘AHS 2015 Value Labels.csv’ is read in as df\_valuelabels and used to create a dictionary for every attribute so that the values can be mapped into df\_household\_clean. After that, the columns of df\_household\_clean are renamed using the new names in df\_colnames.

Finally, the null values in df\_household\_clean, which are stored as -9 or -6, are changed in NaN values. Any record which has a null value in the rating\_house or rating\_neighborhood columns is dropped, as these two columns are necessary for this project.

**4.0 Final DataFrame**

The resulting DataFrame, df\_household\_clean, has 58,233 entries and 129 columns. All of the values are either remapped to the correct category, or are integers.

The data wrangling steps and the data can be viewed here: https://github.com/kellywong1314/Springboard-Data-Science/blob/master/Capstone%201%20-%20American%20Housing%20Satisfaction/Data%20Wrangling\_American%20Housing%20Survey%202015.ipynb

Appendix B

|  |  |  |
| --- | --- | --- |
| **petty\_crime** | 2,642.07 | 0.00E+00 |
| **serious\_crime** | 2,449.85 | 0.00E+00 |
| **rating\_neighborhood** | 2,106.82 | 0.00E+00 |
| **dishwash** | 1,746.21 | 0.00E+00 |
| **wall\_crack** | 1,611.99 | 0.00E+00 |
| **garage** | 1,492.11 | 2.470328e-322 |
| **paint\_peeling** | 1,332.31 | 1.98E-288 |
| **good\_schools** | 1,309.70 | 2.49E-283 |
| **washer** | 1,168.80 | 1.21E-253 |
| **near\_trash** | 1,107.91 | 0.00E+00 |
| **number\_upkeep\_probs** | 1,056.57 | 0.00E+00 |
| **adequacy** | 1,028.01 | 0.00E+00 |
| **floor\_hole** | 986.54 | 9.52E-215 |
| **tenure** | 875.32 | 0.00E+00 |
| **home\_better\_than\_last** | 804.17 | 0.00E+00 |
| **wall\_slope** | 743.26 | 2.99E-162 |
| **roof\_hole** | 730.77 | 1.41E-159 |
| **roof\_sag** | 729.04 | 3.30E-159 |
| **toilet\_broke** | 689.17 | 5.17E-151 |
| **missing\_siding** | 677.60 | 3.29E-148 |
| **in\_water\_leaks** | 611.95 | 2.09E-134 |
| **too\_cold** | 585.86 | 1.31E-252 |
| **out\_water\_leaks** | 513.66 | 3.14E-113 |
| **windows\_broken** | 500.06 | 4.02E-110 |
| **porch** | 484.33 | 6.68E-107 |
| **roach** | 443.47 | 0.00E+00 |
| **hoa** | 438.23 | 6.04E-97 |
| **foundation\_crumb** | 421.03 | 4.15E-93 |
| **near\_abandoned** | 417.99 | 1.13E-268 |
| **missing\_shingle** | 409.64 | 1.19E-90 |
| **nh\_better\_than\_last** | 372.62 | 3.63E-233 |
| **dryer** | 349.90 | 3.24E-298 |
| **risk\_of\_flood** | 348.58 | 1.46E-77 |
| **windows\_boarded** | 348.32 | 2.01E-77 |
| **fireplace** | 348.10 | 4.49E-224 |
| **no\_running\_water** | 310.31 | 2.84E-69 |
| **musty** | 305.53 | 6.90E-258 |
| **rodent** | 260.25 | 4.56E-222 |
| **near\_bar\_windows** | 230.88 | 1.35E-100 |
| **unit\_size** | 206.45 | 0.00E+00 |
| **num\_bedrooms** | 203.81 | 3.64E-216 |
| **num\_bathrooms** | 203.23 | 0.00E+00 |
| **subdivision** | 199.27 | 3.89E-45 |
| **num\_laundry** | 196.61 | 7.93E-86 |
| **num\_dining** | 183.29 | 4.45E-80 |
| **stairs** | 181.54 | 2.98E-41 |
| **household\_type** | 150.50 | 2.65E-190 |
| **bldg\_type** | 141.71 | 5.91E-266 |
| **unit\_floors** | 131.30 | 1.42E-57 |
| **total\_rooms** | 118.24 | 5.27E-293 |
| **sewerbreakdowns** | 83.95 | 3.46E-88 |
| **entry\_sys** | 82.38 | 1.26E-19 |
| **windows\_barred** | 73.33 | 1.13E-17 |
| **fuse\_blow** | 68.31 | 1.90E-71 |
| **housing\_cost** | 66.37 | 6.33E-214 |
| **num\_kitchens** | 55.28 | 1.06E-13 |
| **fridge** | 54.94 | 1.26E-13 |
| **year\_built** | 54.09 | 2.63E-109 |
| **stairs\_broken** | 50.14 | 1.50E-12 |
| **near\_transit** | 49.54 | 1.97E-12 |
| **is\_condo** | 21.27 | 3.99E-06 |
| **hud\_subsidized** | 20.95 | 8.15E-10 |
| **partner\_household** | 12.53 | 3.47E-10 |
| **rent\_subsidy** | 11.65 | 7.06E-15 |
| **num\_people** | 10.86 | 6.91E-27 |
| **interview\_lang** | 7.75 | 4.29E-04 |
| **stories** | 6.82 | 3.02E-07 |
| **hh\_age** | 5.18 | 1.04E-38 |
| **metro\_area** | 5.01 | 1.63E-10 |
| **year\_moved\_in** | 5.00 | 2.07E-45 |
| **gut\_rehab** | 4.41 | 3.57E-02 |
| **manager\_onsite** | 3.79 | 9.88E-03 |
| **monthly\_rent** | 3.59 | 6.80E-43 |
| **rent\_control** | 3.12 | 7.76E-02 |
| **bath\_exclusive** | 0.15 | 6.97E-01 |
| **kitchen\_exclusive** | 0.00 | 9.55E-01 |