Housing Project Report

David Lieberman, Lisa Lin, Quan Le

2024-10-17

Introduction

This report details a case study of the augmentation of a particular real estate dataset, and the exploration of said dataset. We describe first the real estate data, and then the augmenting of the data with features derived from a Hartford crime dataset. We provide some explorations of the crime data, and construct a collection of models that hope to explain some of the variation in the real estate data.

Data

CT Real Estate

The Connecticut Real Estate data is a collection of data on properties across Connecticut. It includes, among other things, houses, condos, businesses, and apartments. The data includes the geometric shapes of each land parcel, and other features like the number of bedrooms, bathrooms, its appraised price, its address, and so on. We defer the reader to the many other reports that will also concern this dataset.

Hartford Crime Incidents

The data can be obtained from either arcgis or Open Data Hartford. It consists of incident-level data on crimes in Hartford, CT. Of particular interest to us is the fact that each incident includes geographic data of where the incident occurred, and includes incidents from 2005 to 2021. Some relevant features of the crime are:

- Date: the date of the incident in YYYY-MM-DD format,
- Time_24HR: the time of the incident in a 24-hour HHMM format,
- Address: the address of the incident,
- UCR_1_Category: the Uniform Crime Reporting category of the incident (e.g. "SUICIDE", "HOMICIDE", "LARCENY", "WEAPONS OFFENSES"),
- UCR_1_Description: the UCR description of the incident with more detailed information than the category (e.g. "ABANDON-FAMILY", "LARC1-BICYCLE"),
- geometry: the GIS point data for each incident.

The other features are less relevant. There are case numbers, neighborhoods, and partial reporting for UCR2 categories.

While there are a number of ways to augment the parcel data, this case study is focused on the two simplest cases – by counting the crime incidents in a certain radius from the real estate parcels, and using kernel density estimation to estimate a crime density. Further analysis can consider more elaborate choices of kernel and bandwidth.

Exploration

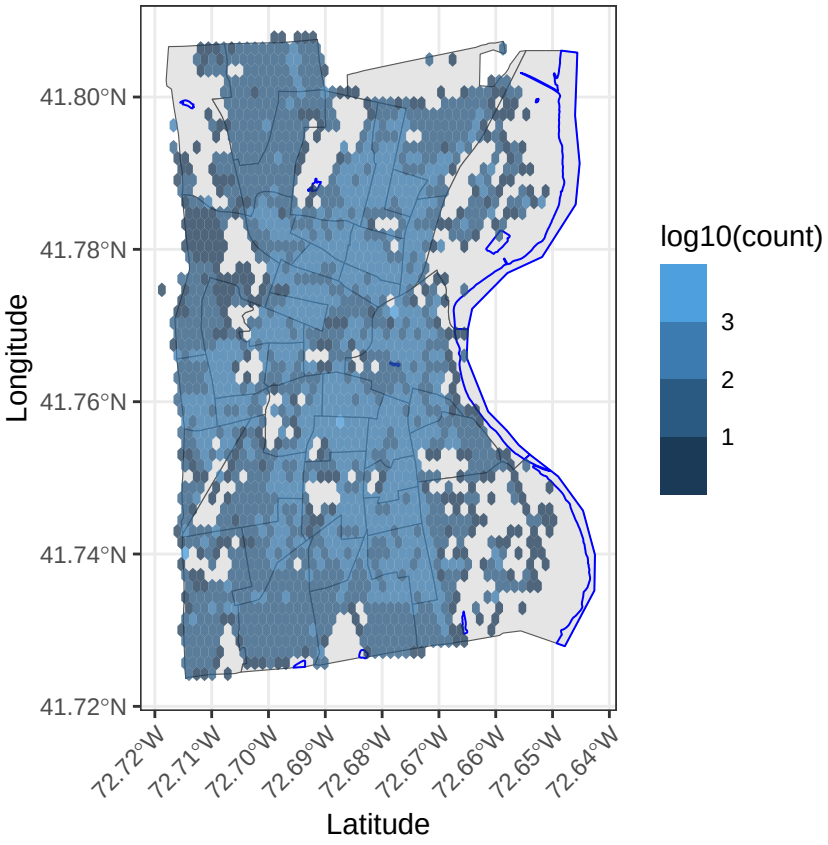
For tractability, we only consider crime data from 2016-2021.

Interactive Map: Kepler

Using Kepler (an interactive data visualization tool), we created a web-available map of both the parcel and crime data, available here. Note that it does take a while (at least a minute) to load, so please be patient.

Crime Heatmap

Below, we can see a heatmap of the log-crime counts from 2016 to 2021. Note the noticeable lack of crimes near the river (outlined in blue), save for three tiles. These can be attributed to bridges. Other gaps in the heatmap can be identified by parks, industrial districts, private universities, or similar areas without much human presence. A side-by-side comparison with the Kepler map usually provides some insight into the particulars.

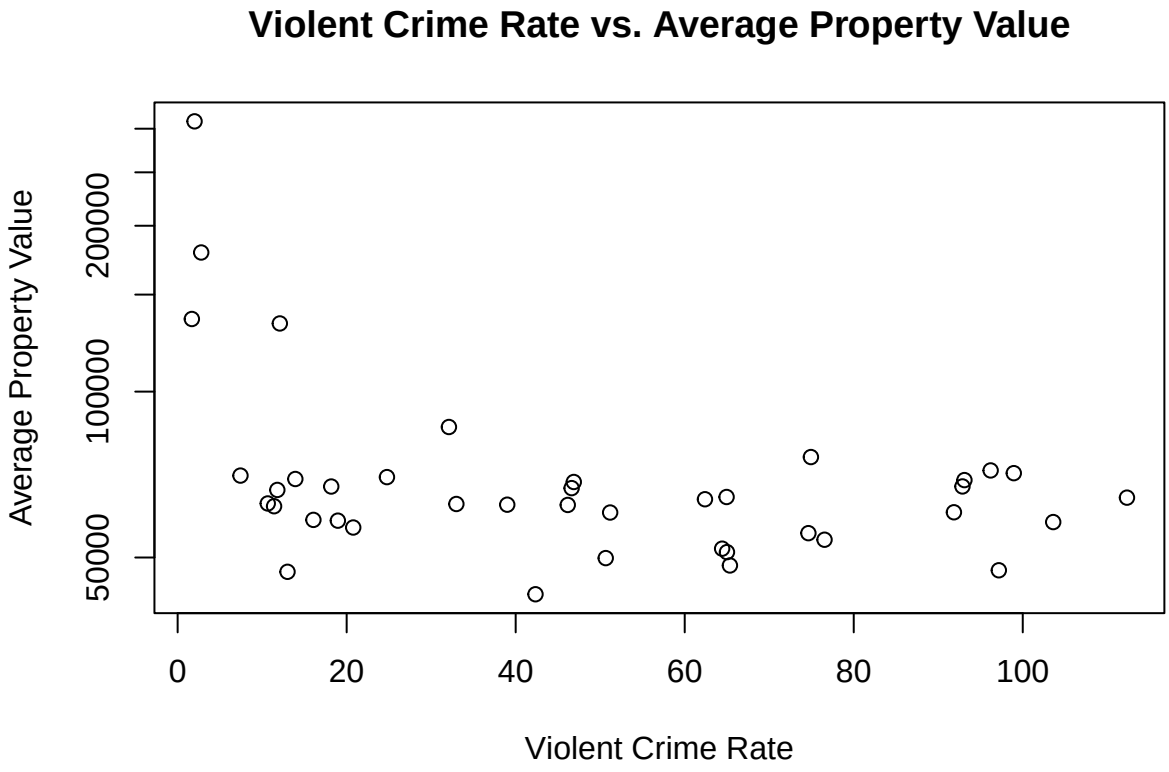


Interactive Crime and Property Value Map

A sample of the data is plotted. Crime is in red, parcels are in black, and the density of crime is in blue. Additional census tract level information like median income and population are included in the tooltip. Interact with the map for details. Note that it will not be visible in the PDF version of this report.

Census Tract Level Aggregated Crime Rate and Average Property Value

Violent crime rates are computed as the average number of violent crimes within 150 meters of a parcel in each tract. The average property value is computed as the mean assessed total value of the parcels in the tract. More details as to what constitutes a "violent crime" can be found in the next section.



Modeling

Preprocessing

Since the data is so large, we prepare some filters and computations in advance and load the processed data. The filters applied to the data are:

- single family homes only, and
- thefts and violent crimes from 2016-2021.

We restrict our attention to single family homes. These are by far the largest group in the dataset. Restricting our analysis to single family homes helps tractability and interpretability, as the dataset includes government and business properties as well. We are also more interested in the relationship between crime and single-family homes in particular, as opposed to all real estate in aggregate.

Since the assessed property values are from 2021, crimes prior to 2016 will be less relevant to the current property values. We also select the crimes that will we predict will be most influential on the home value – violent crime and thefts.

Within the crime data, we filter out major property crimes (robberies, burglaries, larceny, or theft) and violent crimes (assaults, homicides, shootings).

The thefts and violent crimes data is then aggregated spatially in two ways:

1. Count the number of crimes within 150 meters of each parcel
2. Compute kernel density estimation of crimes within a bandwidth of each parcel, where bandwidth is selected via cross-validation

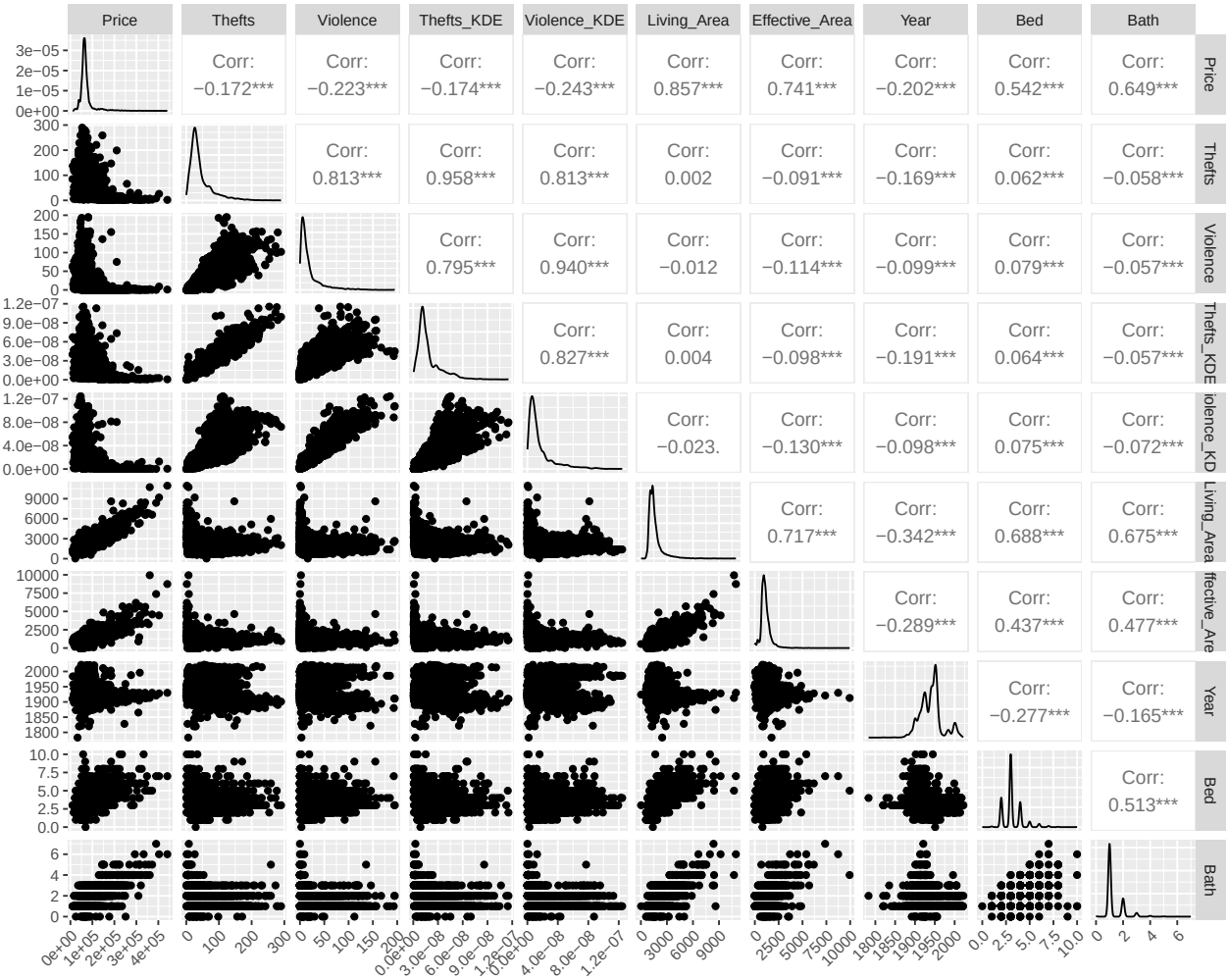
Visualization

The manually curated variables are:

- `Price`: Assessed total value of the parcel
- `Thefts`: Number of thefts (robberies, burglaries, larceny, theft, or stolen property) within 150 meters of the parcel
- `Violence`: Number of violent crimes (assaults, homicides, shootings) within 150 meters of the parcel
- `Thefts_KDE`: Kernel density estimation of thefts
- `Violence_KDE`: Kernel density estimation of violence
- `Living_Area`: Living area of the parcel
- `Effective_Area`: Effective area of the parcel
- `Year`: Approximate year the parcel was built
- `Bed`: Number of bedrooms in the parcel
- `Bath`: Number of bathrooms in the parcel
- `Condition_Description`: The condition of the parcel, one of "Dilapidated", "Very Poor", "Poor", "Fair", "Fair-Avg", "Average", "Avg-Good", "Good", "Good-VG", "Very Good", "Excellent"

Pairs

We can generate the pairwise plots of the variables – scatter plots, densities, and correlations.



The highly correlated numeric covariates are

- Thefts and Violence
- Living Area and Effective Area
- Bed and Bath

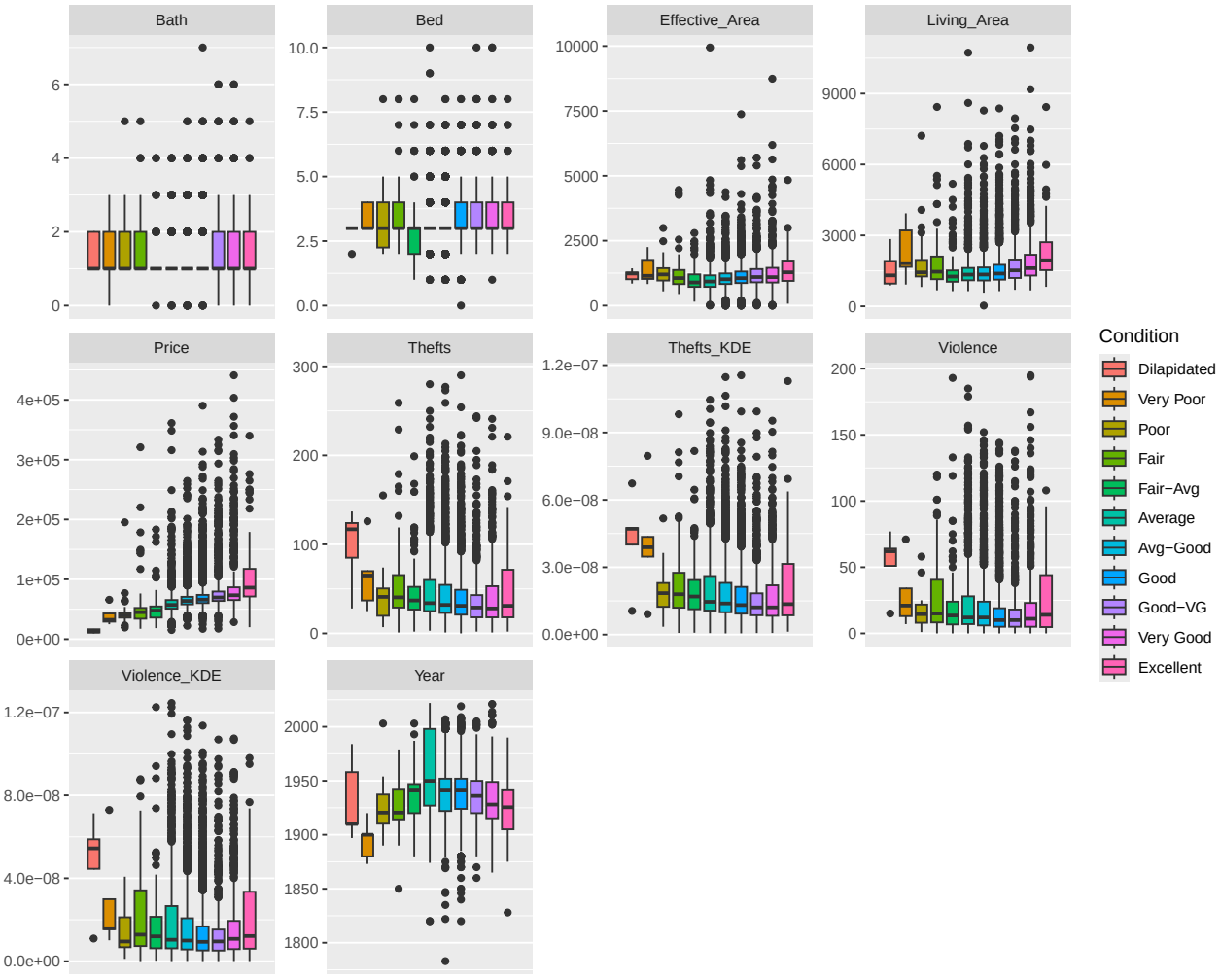
These are generally to be expected. It is not surprising that neighborhoods experiencing high amounts of violent crime would alse experience a high number of thefts. A potential counterargument is that thefts are more likely to occur at wealthier neighborhoods, and these neighborhoods would be less likely to have as much violent crime. There are issues with this argument. Namely, wealthier neighborhoods may have more security, counteracting the effect of a more enticing theft target. In addition, it is unclear as to whether or not this phenomenon would even have a significant effect on the correlation: cases of this kind may be significantly less in number than overall theft.

Living area and effective area being correlated is also unsurprising. The effective area includes the living area. Similarly, it'd be surprising to see a negative correlation between beds and baths.

All covariates appear to correlate with the response variable, `Assessed_Total`.

Condition Factors

Further, by splitting up the conditions, we can see how the numeric variables correlate to these factors.



A couple numeric covariates appear to be correlated with the condition of the parcel:

- Thefts and Violence. There appear to be more thefts and violent crimes near single family homes in `dilapidated` condition. The median thefts near parcels in very poor condition calculated by KDE is relatively higher than the median calculated by counts.
- Year. There appears to be a quadratic relationship between the year a parcel was built and its condition. The median years that parcels were built for those in dilapidated and very poor condition tend to be

earlier than parcels in better condition, suggesting that older homes are more likely to be in worse condition. However, there are also older homes in good condition, and homes in average condition have the most recent median year built.

Analysis

1. Crime counts

We first consider the `Thefts` and `Violence` variables computed by counting crime occurrences within a 150-meter radius of a parcel.

Variable Selection. Given the relatively small number of variables, we forego an automated method (e.g. lasso, stepwise information criteria) and instead inspect our variables for collinearity. We leave out `Thefts` since it is highly correlated with `Violence` but has a lower correlation with `Price` than `Violence` does. For similar reasons, we leave out `Effective_Area` and keep `Living_Area`. We keep both `Bed` and `Bath` for now.

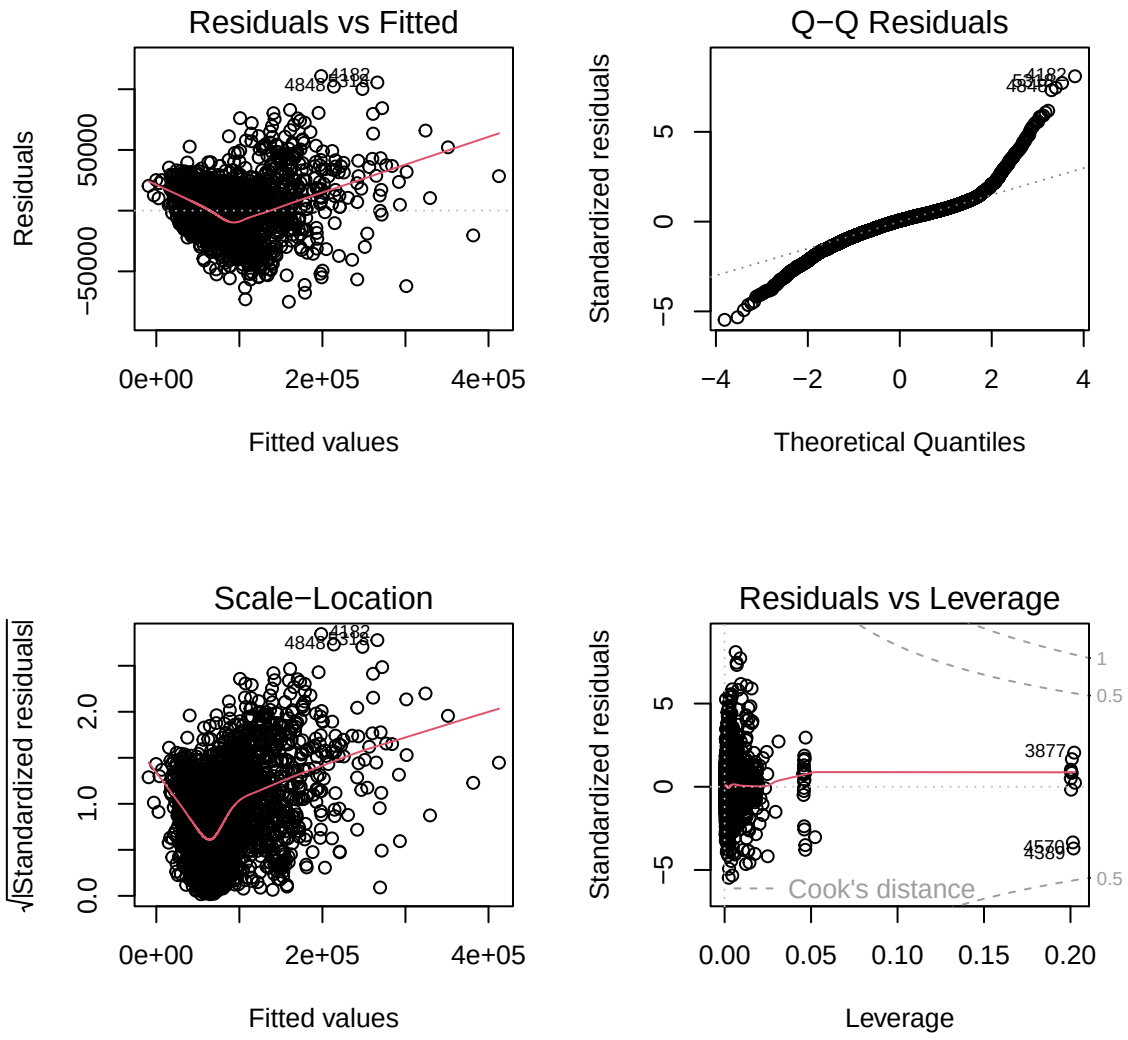
```
##
## Call:
## lm(formula = Price ~ Violence + Living_Area + Year + Bed + Bath +
##     Condition, data = X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75062  -7023    856   6819 110854
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.458e+05  1.440e+04 -17.063  < 2e-16 ***
## Violence           -2.521e+02  6.990e+00 -36.065  < 2e-16 ***
## Living_Area         3.349e+01  3.255e-01 102.893  < 2e-16 ***
## Year                1.132e+02  6.656e+00  17.002  < 2e-16 ***
## Bed                -2.306e+03  2.293e+02 -10.061  < 2e-16 ***
## Bath                5.426e+03  3.556e+02  15.258  < 2e-16 ***
## ConditionVery Poor  6.029e+02  8.696e+03   0.069   0.9447
## ConditionPoor       1.345e+04  6.813e+03   1.974   0.0484 *
## ConditionFair       2.685e+04  6.408e+03   4.190 2.82e-05 ***
## ConditionFair-Avg   3.280e+04  6.258e+03   5.241 1.65e-07 ***
## ConditionAverage    4.180e+04  6.163e+03   6.784 1.27e-11 ***
## ConditionAvg-Good   4.936e+04  6.158e+03   8.017 1.26e-15 ***
## ConditionGood       5.025e+04  6.157e+03   8.161 3.90e-16 ***
## ConditionGood-VG    5.442e+04  6.176e+03   8.810  < 2e-16 ***
## ConditionVery Good  5.704e+04  6.179e+03   9.231  < 2e-16 ***
## ConditionExcellent  6.455e+04  6.330e+03  10.197  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13740 on 7204 degrees of freedom
## Multiple R-squared:  0.8252, Adjusted R-squared:  0.8248
## F-statistic:  2267 on 15 and 7204 DF,  p-value: < 2.2e-16
```

Suppose for a moment that the data satisfy the linear regression assumptions. We will analyze the diagnostic plots in the next subsection.

Even with a relatively naive model, we note that we have a fairly good R-squared value at around 0.8 or so. Regarding the coefficients, we see that the coefficient for violence is negative, and comes to about -250 dollars per violent crime. The bed and bath coefficients have opposite sign. As hoped, the condition factors (with `Dilapidated` as baseline) are ordered in the expected manner.

Each of the numeric variables has a significant p-value, smaller than R's machine precision. Most of the condition factors are also significant, save for `Very Poor`, whereas `Poor` is only a bit significant. We can generally interpret these to mean that between sub-fair conditions, the differences between house price are not so significant. A potential reason for this is the fact that houses in poorer condition have little actual value derived from the structure itself – assessors might be considering these properties as requiring a significant renovation or even a teardown.

The largest t-statistics in magnitude are, in decreasing order, `Living_Area`, `Violence`, and `Year`.



Diagnostics

The residuals cast doubt on our linear regression assumptions. In particular, it seems that the residuals are neither independent, homoscedastic, nor normal. The q-q plot has a heavy tails, and we can see non-constant trendlines in the residual plots.

Transformations. From the correlation plots, we can see that `Price`, `Living_Area`, and `Violence` are right-skewed. To adjust for this, we take the log of `Price` and the square root of `Living_Area` and `Violence`.

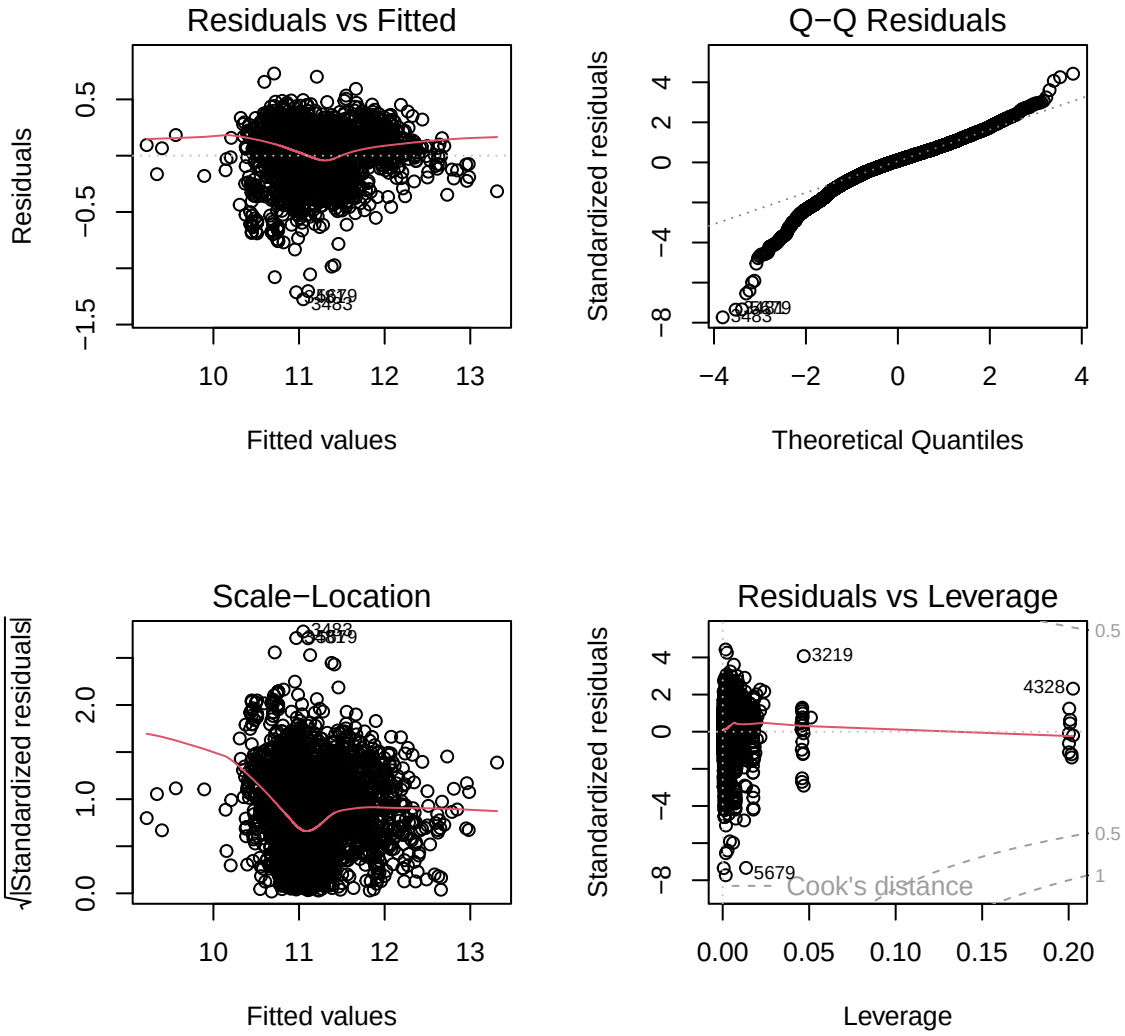
```
##
## Call:
## lm(formula = log(Price) ~ sqrt(Violence) + sqrt(Living_Area) +
##     Year + Bed + Bath + Condition, data = X)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.27511 -0.07695  0.01528  0.09805  0.72980
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.884e+00  1.771e-01  38.860  < 2e-16 ***
## sqrt(Violence)      -4.028e-02  9.175e-04 -43.908  < 2e-16 ***
## sqrt(Living_Area)    2.823e-02  3.803e-04  74.234  < 2e-16 ***
## Year                 9.114e-04  8.078e-05  11.282  < 2e-16 ***
## Bed                 -9.644e-03  2.818e-03  -3.422 0.000626 ***
## Bath                 4.061e-02  4.161e-03   9.759  < 2e-16 ***
## ConditionVery Poor   7.605e-01  1.045e-01   7.279 3.71e-13 ***
## ConditionPoor        9.360e-01  8.185e-02  11.436  < 2e-16 ***
## ConditionFair        1.068e+00  7.699e-02  13.875  < 2e-16 ***
## ConditionFair-Avg    1.131e+00  7.519e-02  15.043  < 2e-16 ***
## ConditionAverage     1.377e+00  7.404e-02  18.603  < 2e-16 ***
## ConditionAvg-Good    1.500e+00  7.398e-02  20.271  < 2e-16 ***
## ConditionGood        1.517e+00  7.398e-02  20.507  < 2e-16 ***
## ConditionGood-VG     1.552e+00  7.421e-02  20.911  < 2e-16 ***
## ConditionVery Good   1.580e+00  7.424e-02  21.277  < 2e-16 ***
## ConditionExcellent   1.650e+00  7.605e-02  21.702  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1651 on 7204 degrees of freedom
## Multiple R-squared:  0.7622, Adjusted R-squared:  0.7617
## F-statistic:  1540 on 15 and 7204 DF,  p-value: < 2.2e-16
```

If the linear regression assumptions hold, then not much changes in the interpretation of the coefficients between models – the bath and bed coefficients have opposite sign, violence is negative, and the condition factors are ordered as expected.

A notable difference is the significances – now all the p-values are quite small.

The largest t-statistics in magnitude are `sqrt(Living_Area)` and `sqrt(Violence)`, in decreasing order.



Diagnostics These seem better, but there is a group of notable residuals. We see that this cluster of properties is priced well below what our model expects. We explore this discrepancy in the subsequent section.

Otherwise, the plots are ok. It seems that for the most part, the residuals are independent and homoscedastic. The q-q plot is also better, but still has a deviation from the normal quantiles at the tails.

The residual plots look much better for the transformed model.

Outliers There are 31 parcels beyond 4 standard errors below the regression line. This means the model is severely overestimating the value of these parcels. Let's find out what these parcels are.

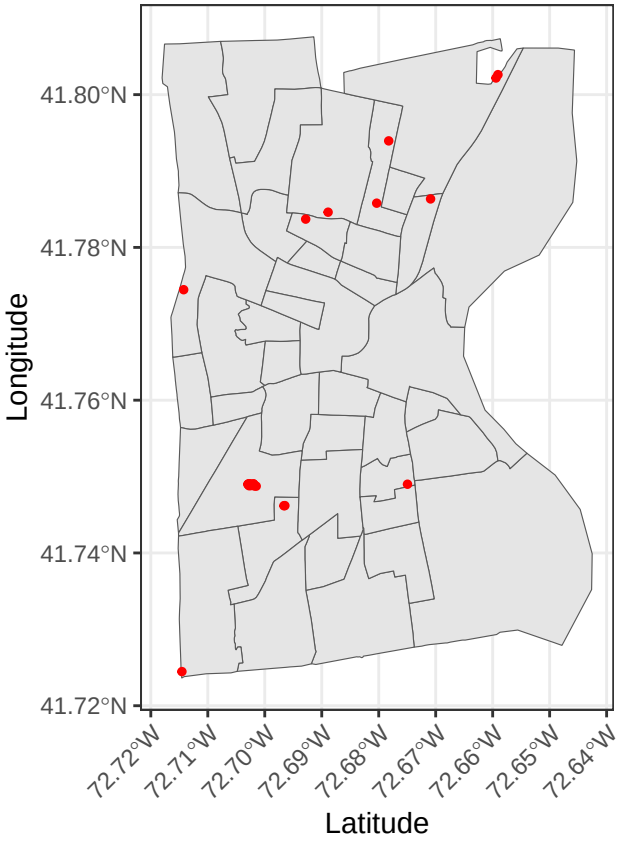
- Geography: There is a cluster of 19 offending parcels near (-72.7, 41.75) in Census Tract 5049. The remaining parcels are scattered throughout Hartford. Interact with the map to see the details of each parcel.
- Condition: The "large" residuals occur for parcels in better than `Average` condition, although the majority of them are `Average` or worse.
- Numeric covariates: The residuals are highly correlated with the living area and number of bedrooms.

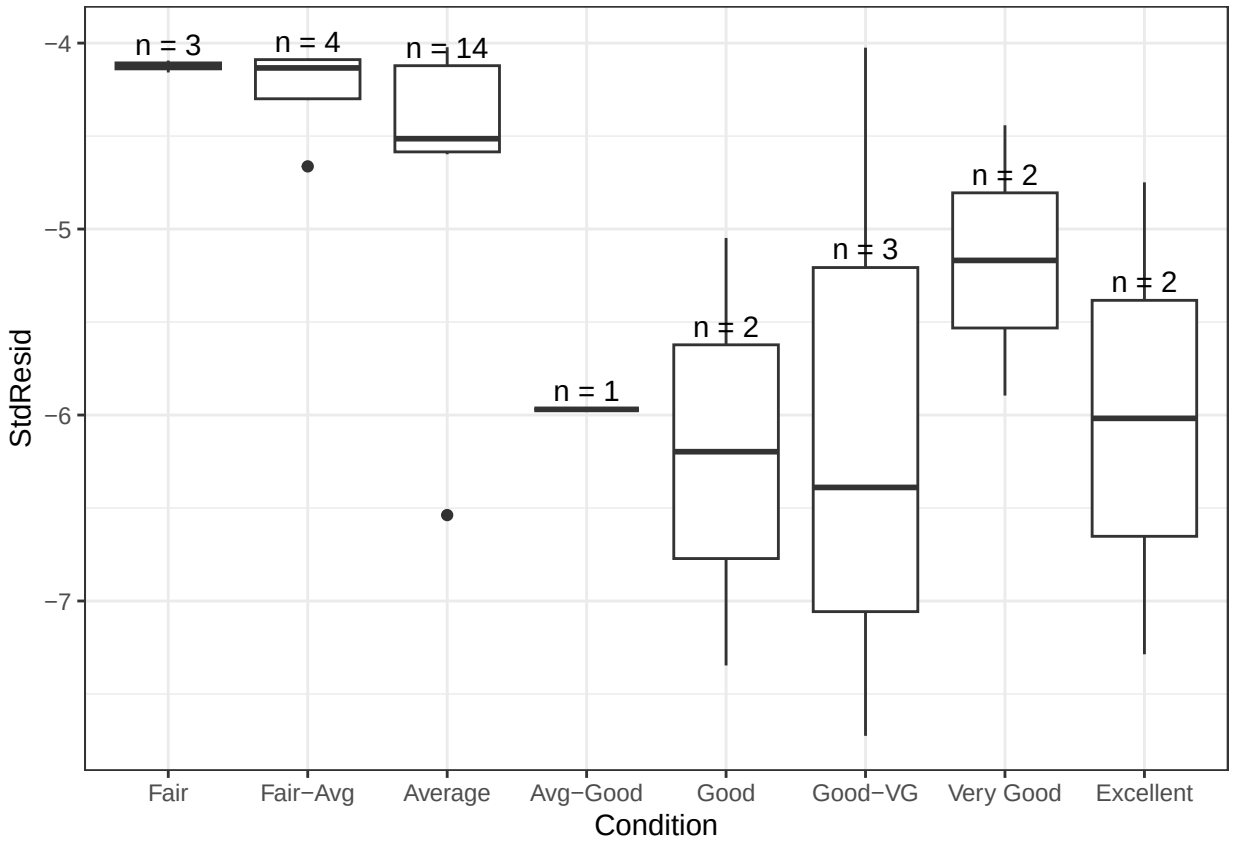
Solutions. We try the following fixes:

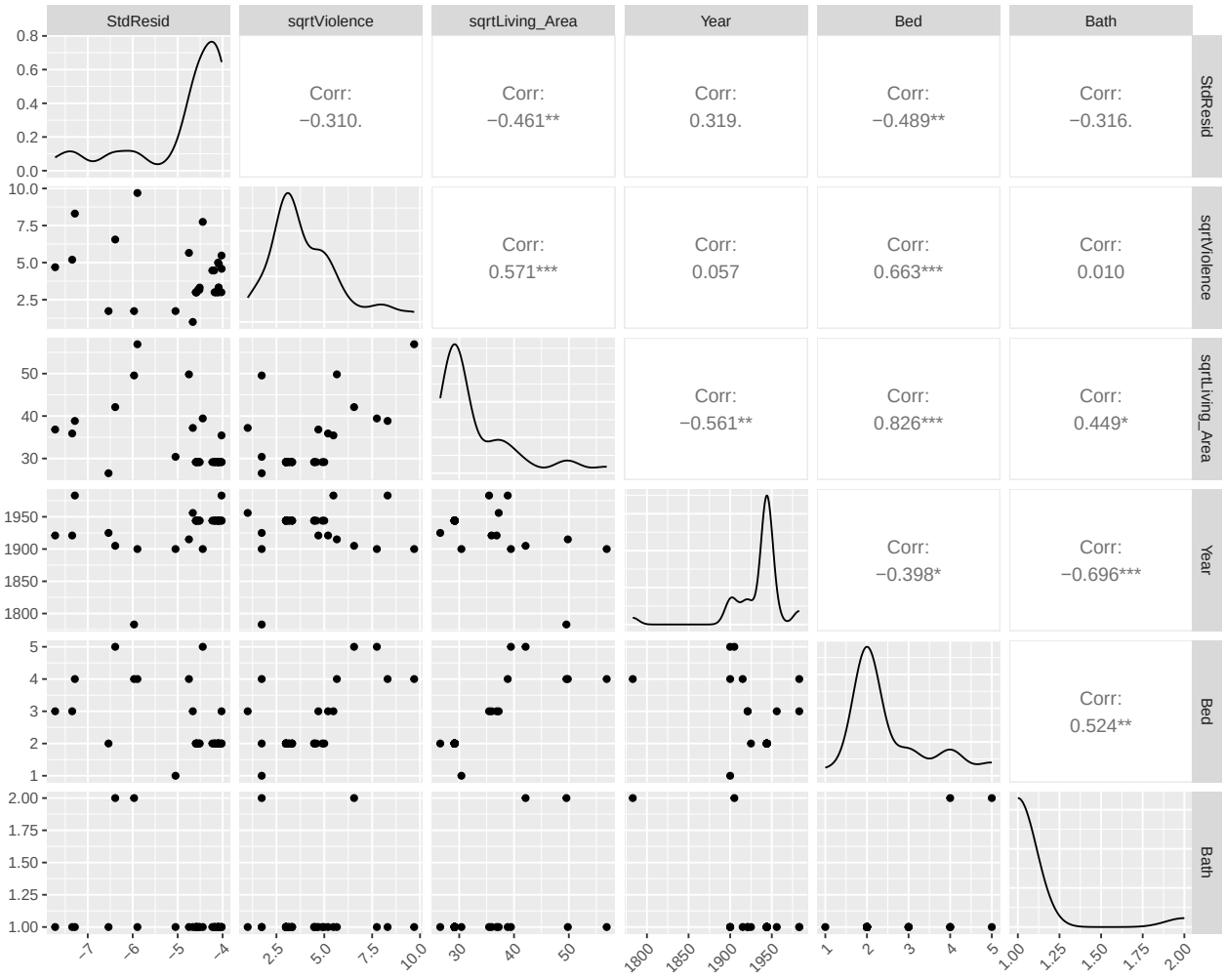
- Log transform living area: Refer to the previous correlation plots and observe that Price and Living Area appear to have a strongly correlated linear correlation. However, we performed a `log` transformation

on the Price variable but a square-root transformation on the Living Area variable. This asymmetry may be the cause of the large residuals.

- Drop number of bedrooms: The number of bedrooms is highly correlated with the residuals, and we might infer that most of the information provided by this variable is already captured in the number of bathrooms.
- Drop the 19 parcels: These parcels are in a small geographic area and may be subject to some unobserved spatial effect.




n = 3
n = 4
n = 14
n = 1
n = 2
n = 3
n = 2
n = 2
−4
−5
−6
−7
StdResid
Fair
Fair−Avg
Average
Avg−Good
Good
Good−VG
Very Good
Excellent
Condition

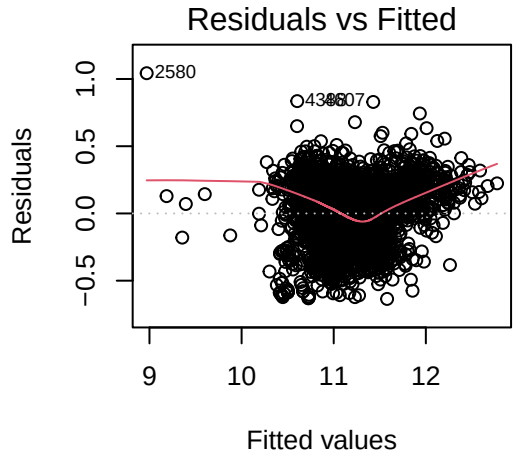


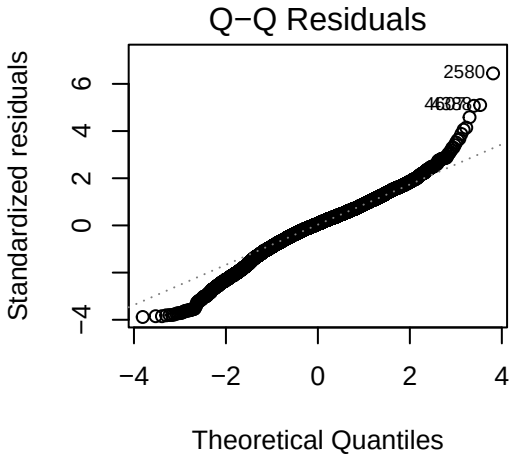
Let's test our hypotheses (See Demo 2 for full summary output and diagnostic plots).

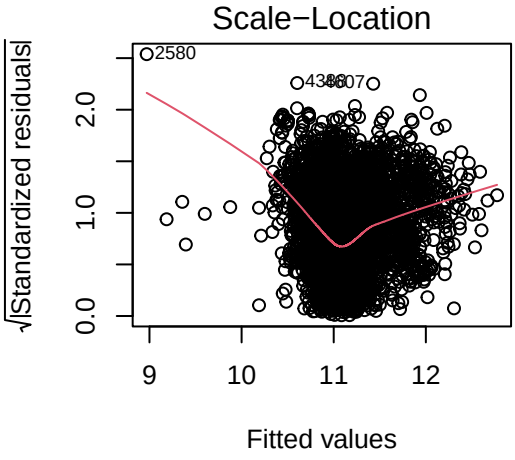
- After taking the log of `Living_Area`, the residuals are more symmetrically distributed. We may have sacrificed some upper tail normality for the lower tail, since there appear to be more residuals larger than 4 now. However, we have reduced the total number of residuals beyond 4 SE's from 33 to 21. Importantly, the `Beds` variable is no longer statistically significant.
- Dropping the number of bedrooms has no noticeable effect on the model.
- Dropping the 19 parcels increases the adjusted R^2 from 0.75 to 0.76 and further decreases the number of residuals beyond 4 SE's to 5. Thus, these parcels represented a significant source of error in the model and might be a good point of further investigation. The diagnostic plots for this model are printed below.

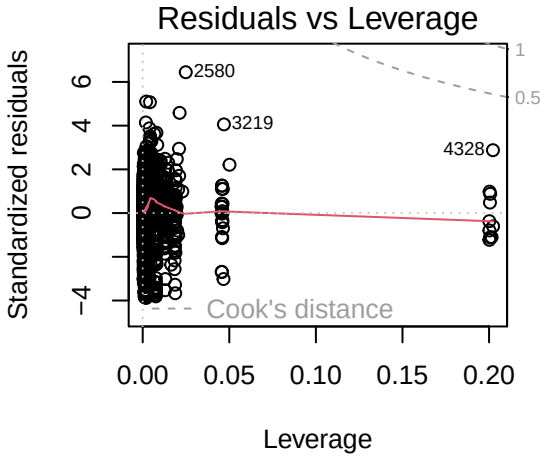
```
##
## Call:
## lm(formula = log(Price) ~ sqrt(Violence) + log(Living_Area) +
##     Year + Bath + Condition, data = XO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63568 -0.08865  0.01138  0.09956  1.04299
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          4.468e+00  1.920e-01  23.274  < 2e-16 ***
## sqrt(Violence)      -4.375e-02  9.059e-04 -48.297  < 2e-16 ***
## log(Living_Area)     5.430e-01  6.697e-03  81.076  < 2e-16 ***
## Year                 6.595e-04  8.042e-05   8.202 2.78e-16 ***
## Bath                 7.561e-02  3.916e-03  19.311  < 2e-16 ***
## ConditionVery Poor   7.680e-01  1.037e-01   7.403 1.49e-13 ***
## ConditionPoor        9.317e-01  8.127e-02  11.464  < 2e-16 ***
## ConditionFair        1.104e+00  7.661e-02  14.406  < 2e-16 ***
## ConditionFair-Avg    1.144e+00  7.470e-02  15.314  < 2e-16 ***
## ConditionAverage     1.375e+00  7.352e-02  18.702  < 2e-16 ***
## ConditionAvg-Good    1.485e+00  7.346e-02  20.215  < 2e-16 ***
## ConditionGood        1.501e+00  7.346e-02  20.434  < 2e-16 ***
## ConditionGood-VG     1.542e+00  7.368e-02  20.931  < 2e-16 ***
## ConditionVery Good   1.572e+00  7.372e-02  21.318  < 2e-16 ***
## ConditionExcellent   1.669e+00  7.557e-02  22.082  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1639 on 7174 degrees of freedom
## Multiple R-squared:  0.755,  Adjusted R-squared:  0.7545
## F-statistic:  1579 on 14 and 7174 DF,  p-value: < 2.2e-16
```





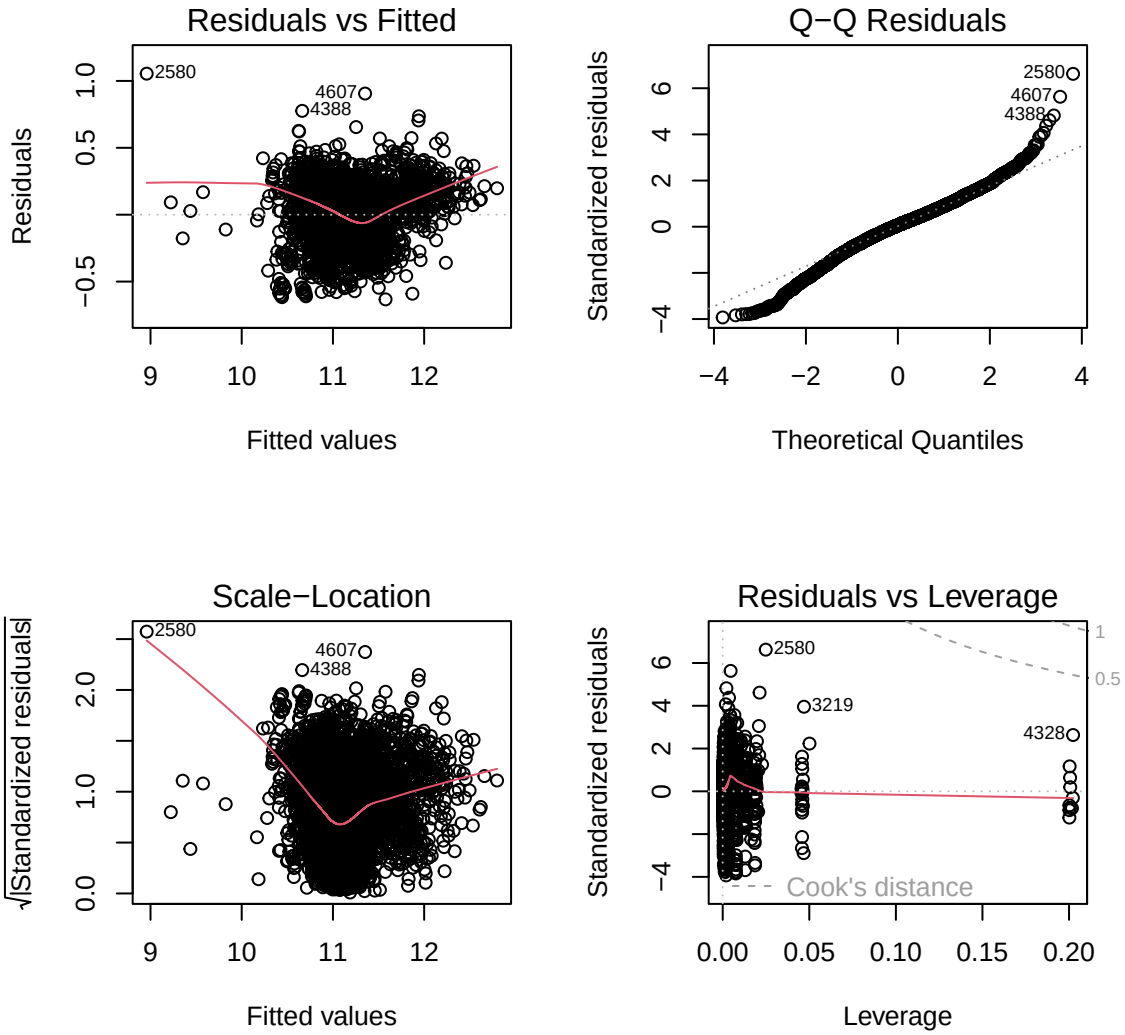




2. Crime KDE

Replacing the `Violence` computed by counts with `Violence_KDE` computed by kernel density estimation increases the adjusted R^2 and reduces residual standard error. From the plots, there is not a substantial change in the residuals. Thus, we can conclude that the KDE of violent crimes is a better predictor of property value than the count of violent crimes. Our final model is:

$$\log \text{Price} = \beta_0 + \beta_1 \sqrt{\text{Violence\_KDE}} + \beta_2 \log(\text{Living\_Area}) + \beta_3 \text{Year} + \beta_4 \text{Bath} + \beta_5 \text{Condition}$$



The diagnostic plots are better in the sense that there is no longer a cluster of large residuals. The trendline in the residuals vs. fitted plot still features that undesireable kink in the middle. The remainder of the plots show little deviation from those seen in previous sections.

Conclusion

Violence, while considering other variables, is a significant and negative predictor of property value. The KDE of violent crimes is a better predictor than the count of violent crimes.

While our model does not exactly line up with the data, it is a good start. We can see that our model explains a significant amount of the variation in the log assessed price.

Further Directions

Our analysis was fairly constrained to ensure that our analysis came to a conclusion. There are a great many different directions that we could take. In terms of variable selection, we could expand our analysis to consider multifamily homes. Alternatively, we could modify the kinds of crimes that we consider. There are dozens and dozens of categories alone, not to mention the specific crime descriptions that belong in each category.

Other feature choices are possible. We could consider different kernels and bandwidths for KDE, or different thresholds for the crime counts. We could add interaction and polynomial terms.

Other transformations beyond the standard log and root that we use here are possible. Other models are possible, especially if we are less interested in modeling variation and more interested in prediction.

Other interpretations are possible. While we can compute t-statistics and p-values, or compare magnitudes of the coefficients, there are other methods that consider feature importance.

We could choose a different problem. Our analysis focused on regressing the assessed value of a property, but another analysis could look for the best places to rob people in Hartford.

References

Spatial regression

https://oerstatistics.wordpress.com/wp-content/uploads/2016/03/intro_to_r.pdf#page=68.08

https://crd230.github.io/lab8.html

Kernel density estimation

https://seeing-statistics.com/issue4/