Modeling House Prices Using Realtor Data

Agam Singh, Renos Zabounidis,
Manav Kulshrestha
and William Munson

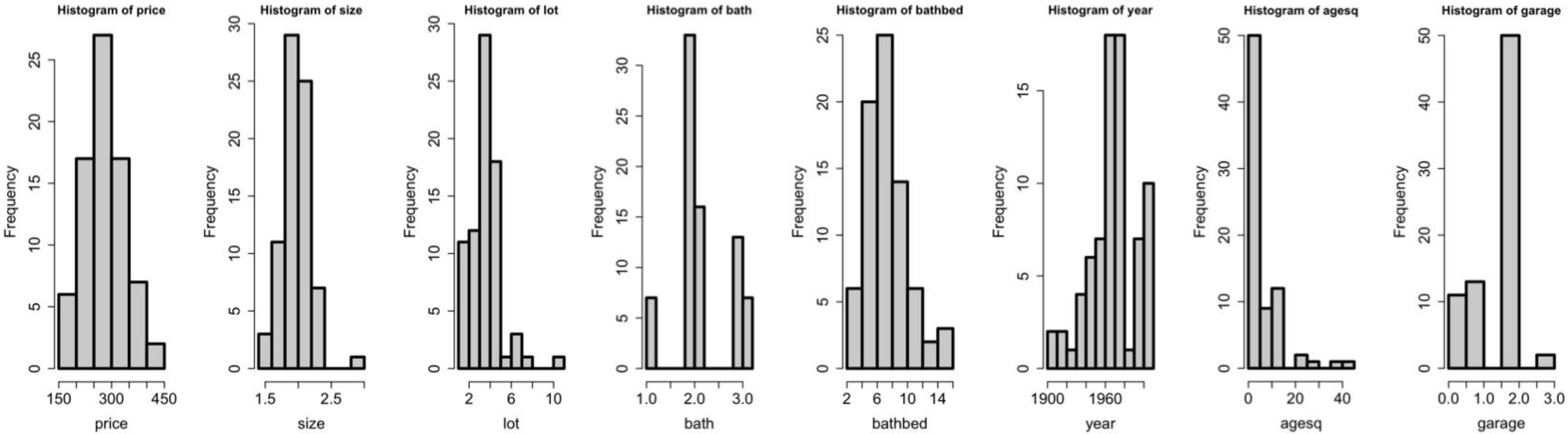
A Little Background Information

- Single-family homes in Eugene, Oregon
- Collected in 2005
- Mix of quantitative and qualitative variables
- 76 data points in total
- 18 variables in total (1 response and 17 predictors)

A Closer Look at the Variables

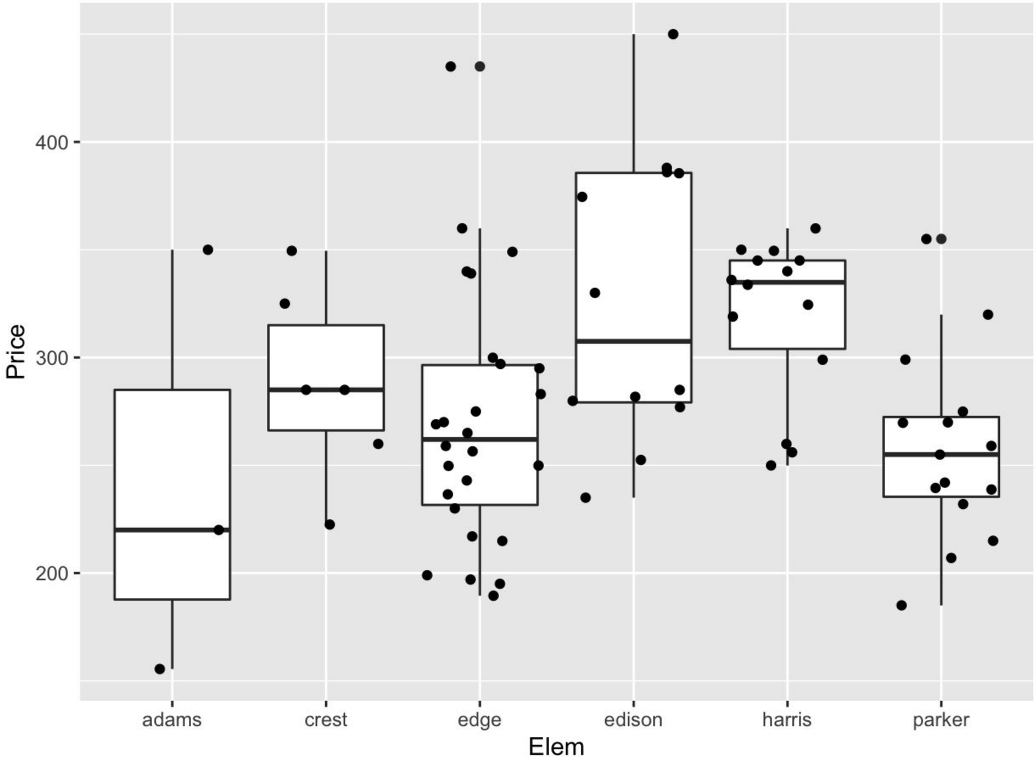
Price	**sale price (thousands of dollars)**
Size	floor size (thousands of square feet)
Lot	lot size category (from 1 to 11)
Bath	number of bathrooms (with half-bathrooms counting as 0.1)
Bed	number of bedrooms (between 2 and 6)
BathBed	interaction of Bath times Bed
Year	year built
Age	age (standardized: (Year-1970)/10)
Agesq	Age squared
Garage	garage size (0, 1, 2, or 3 cars)
Status	act (active listing), pen (pending sale), or sld (sold)
Active	indicator for active listing (reference: pending or sold)
Elem	nearest elementary school (edgewood, edison, harris, adams, crest, or parker)
Edison	indicator for Edison Elementary (reference: Edgewood Elementary)
Harris	indicator for Harris Elementary (reference: Edgewood Elementary)
Adams	indicator for Adams Elementary (reference: Edgewood Elementary)
Crest	indicator for Crest Elementary (reference: Edgewood Elementary)
Parker	indicator for Parker Elementary (reference: Edgewood Elementary)

The Quantitative Variables

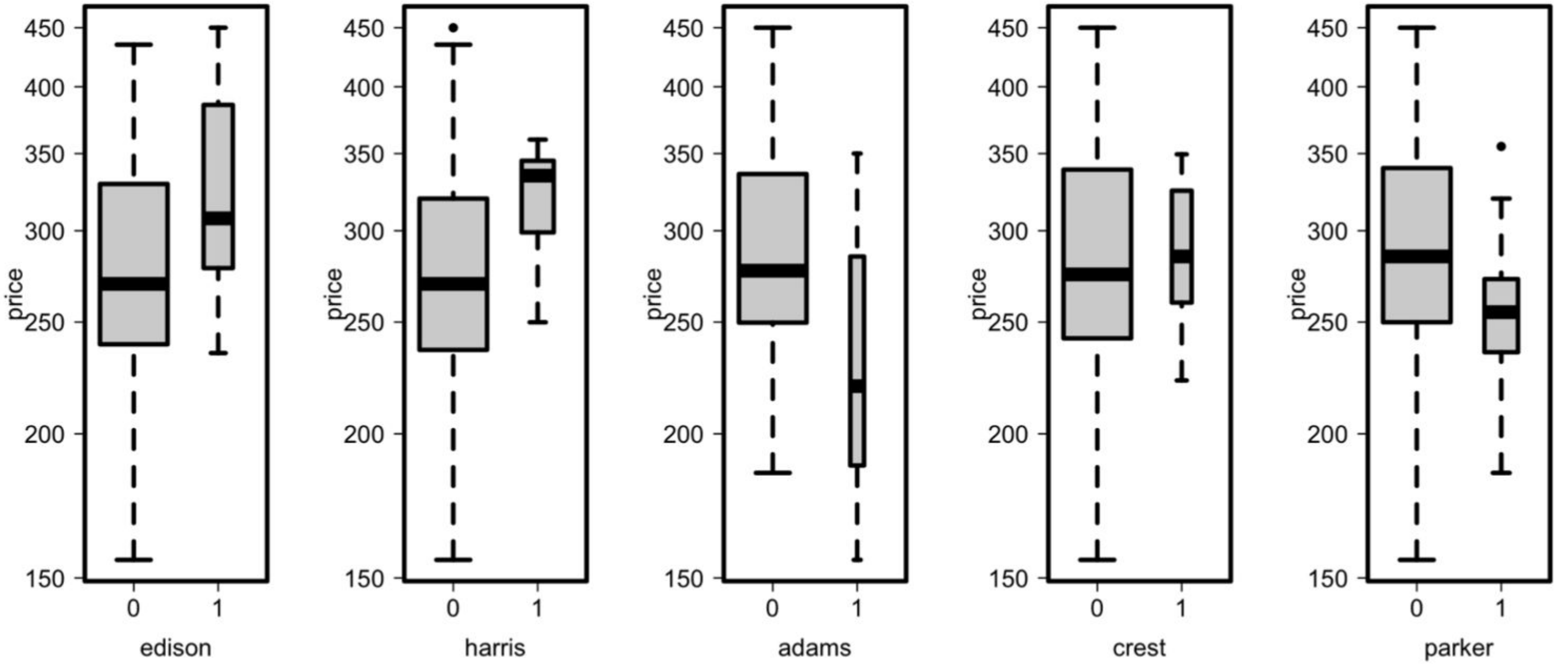


They all seem approximately normal..

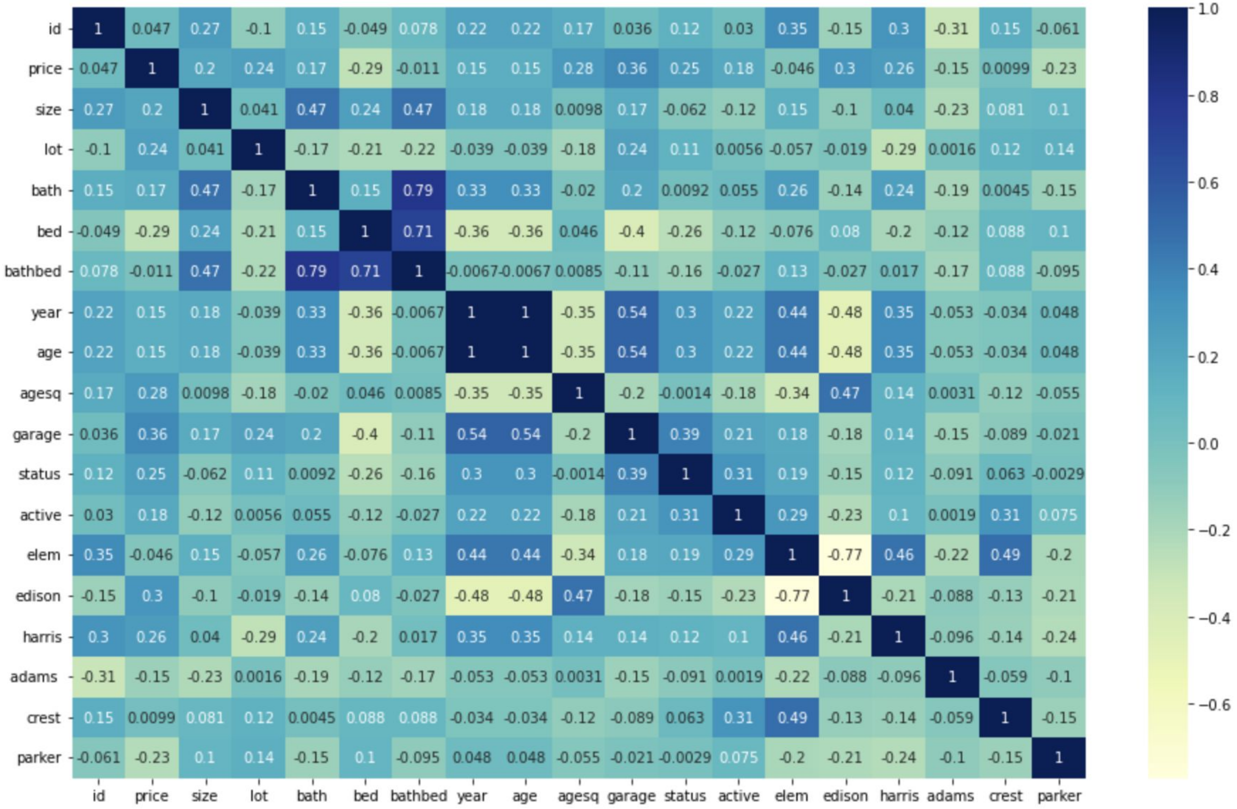
The Schools


400
300
200
Price
adams
crest
edge
edison
harris
parker
Elem

The Schools

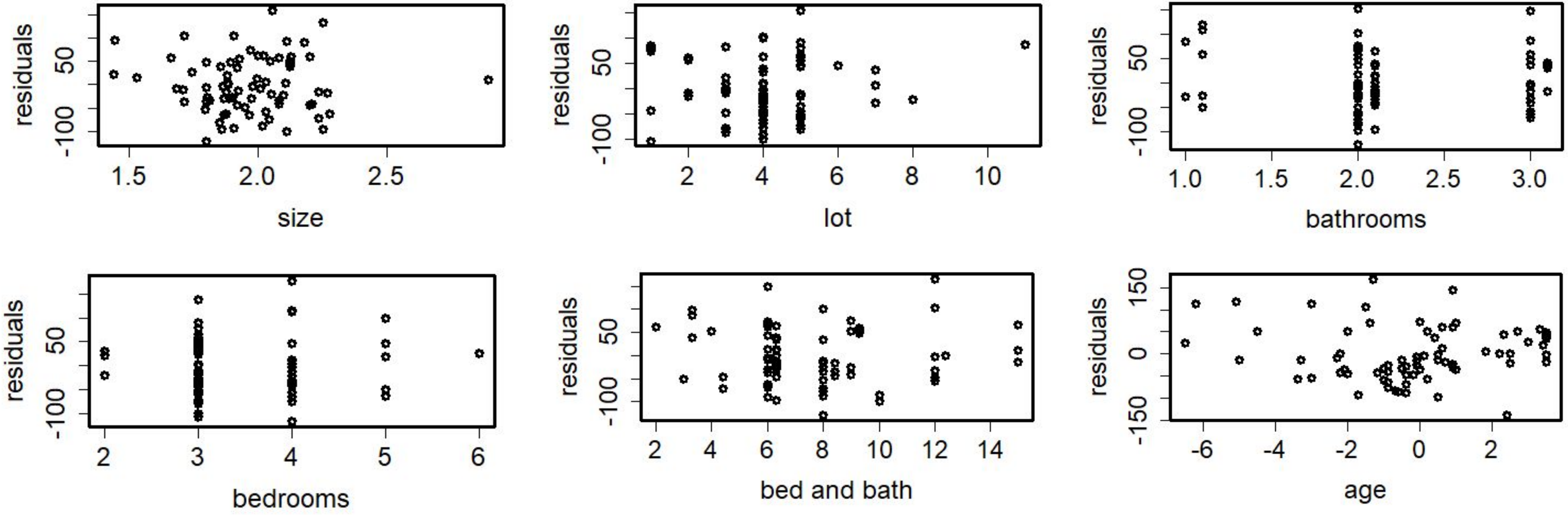

price
450
400
350
300
250
200
150
0
1
edison
harris
adams
crest
parker

Correlation Matrix

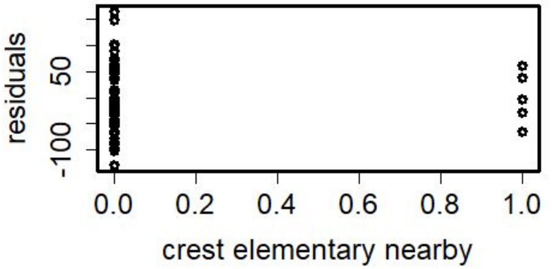


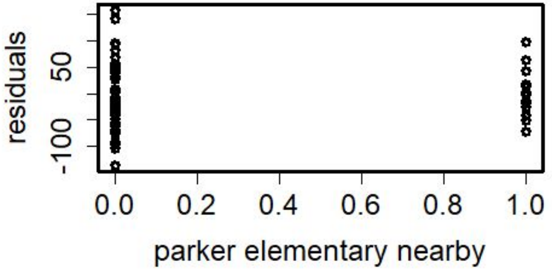
	id	price	size	lot	bath	bed	bathbed	year	age	agesq	garage	status	active	elem	edison	harris	adams	crest	parker
id	1	0.047	0.27	-0.1	0.15	-0.049	0.078	0.22	0.22	0.17	0.036	0.12	0.03	0.35	-0.15	0.3	-0.31	0.15	-0.061
price	0.047	1	0.2	0.24	0.17	-0.29	-0.011	0.15	0.15	0.28	0.36	0.25	0.18	-0.046	0.3	0.26	-0.15	0.0099	-0.23
size	0.27	0.2	1	0.041	0.47	0.24	0.47	0.18	0.18	0.0098	0.17	-0.062	-0.12	0.15	-0.1	0.04	-0.23	0.081	0.1
lot	-0.1	0.24	0.041	1	-0.17	-0.21	-0.22	-0.039	-0.039	-0.18	0.24	0.11	0.0056	-0.057	-0.019	-0.29	0.0016	0.12	0.14
bath	0.15	0.17	0.47	-0.17	1	0.15	0.79	0.33	0.33	-0.02	0.2	0.0092	0.055	0.26	-0.14	0.24	-0.19	0.0045	-0.15
bed	-0.049	-0.29	0.24	-0.21	0.15	1	0.71	-0.36	-0.36	0.046	-0.4	-0.26	-0.12	-0.076	0.08	-0.2	-0.12	0.088	0.1
bathbed	0.078	-0.011	0.47	-0.22	0.79	0.71	1	-0.0067	-0.0067	0.0085	-0.11	-0.16	-0.027	0.13	-0.027	0.017	-0.17	0.088	-0.095
year	0.22	0.15	0.18	-0.039	0.33	-0.36	-0.0067	1	1	-0.35	0.54	0.3	0.22	0.44	-0.48	0.35	-0.053	-0.034	0.048
age	0.22	0.15	0.18	-0.039	0.33	-0.36	-0.0067	1	1	-0.35	0.54	0.3	0.22	0.44	-0.48	0.35	-0.053	-0.034	0.048
agesq	0.17	0.28	0.0098	-0.18	-0.02	0.046	0.0085	-0.35	-0.35	1	-0.2	-0.0014	-0.18	-0.34	0.47	0.14	0.0031	-0.12	-0.055
garage	0.036	0.36	0.17	0.24	0.2	-0.4	-0.11	0.54	0.54	-0.2	1	0.39	0.21	0.18	-0.18	0.14	-0.15	-0.089	-0.021
status	0.12	0.25	-0.062	0.11	0.0092	-0.26	-0.16	0.3	0.3	-0.0014	0.39	1	0.31	0.19	-0.15	0.12	-0.091	0.063	-0.0029
active	0.03	0.18	-0.12	0.0056	0.055	-0.12	-0.027	0.22	0.22	-0.18	0.21	0.31	1	0.29	-0.23	0.1	0.0019	0.31	0.075
elem	0.35	-0.046	0.15	-0.057	0.26	-0.076	0.13	0.44	0.44	-0.34	0.18	0.19	0.29	1	-0.77	0.46	-0.22	0.49	-0.2
edison	-0.15	0.3	-0.1	-0.019	-0.14	0.08	-0.027	-0.48	-0.48	0.47	-0.18	-0.15	-0.23	-0.77	1	-0.21	-0.088	-0.13	-0.21
harris	0.3	0.26	0.04	-0.29	0.24	-0.2	0.017	0.35	0.35	0.14	0.14	0.12	0.1	0.46	-0.21	1	-0.096	-0.14	-0.24
adams	-0.31	-0.15	-0.23	0.0016	-0.19	-0.12	-0.17	-0.053	-0.053	0.0031	-0.15	-0.091	0.0019	-0.22	-0.088	-0.096	1	-0.059	-0.1
crest	0.15	0.0099	0.081	0.12	0.0045	0.088	0.088	-0.034	-0.034	-0.12	-0.089	0.063	0.31	0.49	-0.13	-0.14	-0.059	1	-0.15
parker	-0.061	-0.23	0.1	0.14	-0.15	0.1	-0.095	0.048	0.048	-0.055	-0.021	-0.0029	0.075	-0.2	-0.21	-0.24	-0.1	-0.15	1

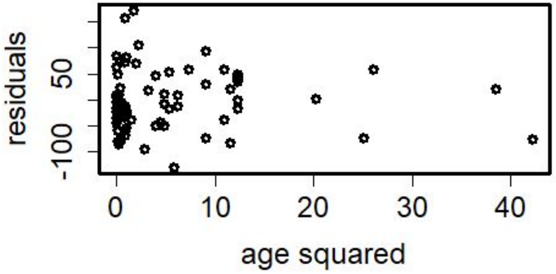
Residual Plots

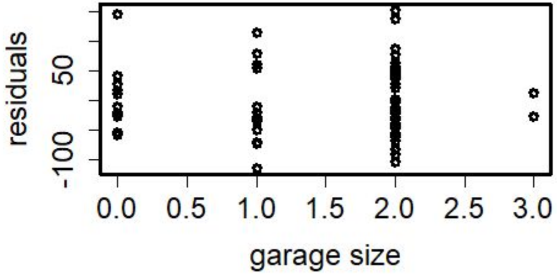


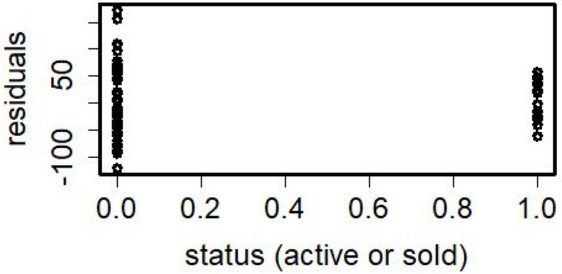
Residual Plots

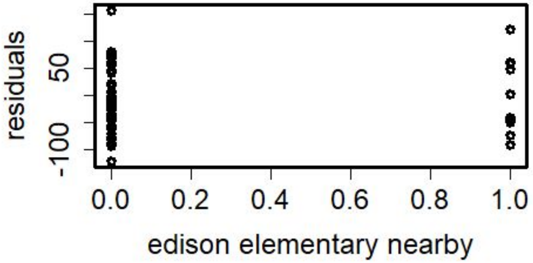


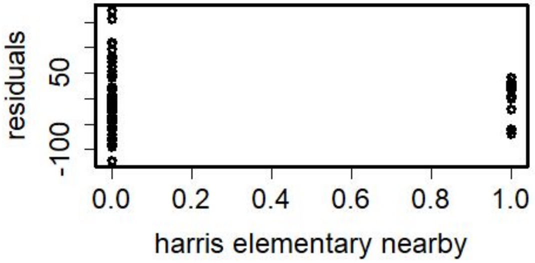


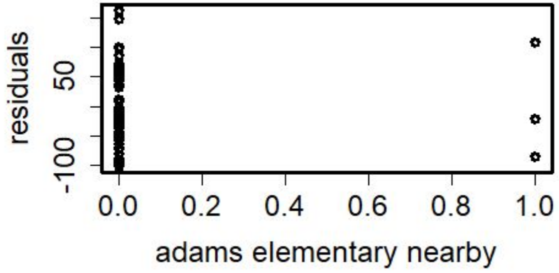




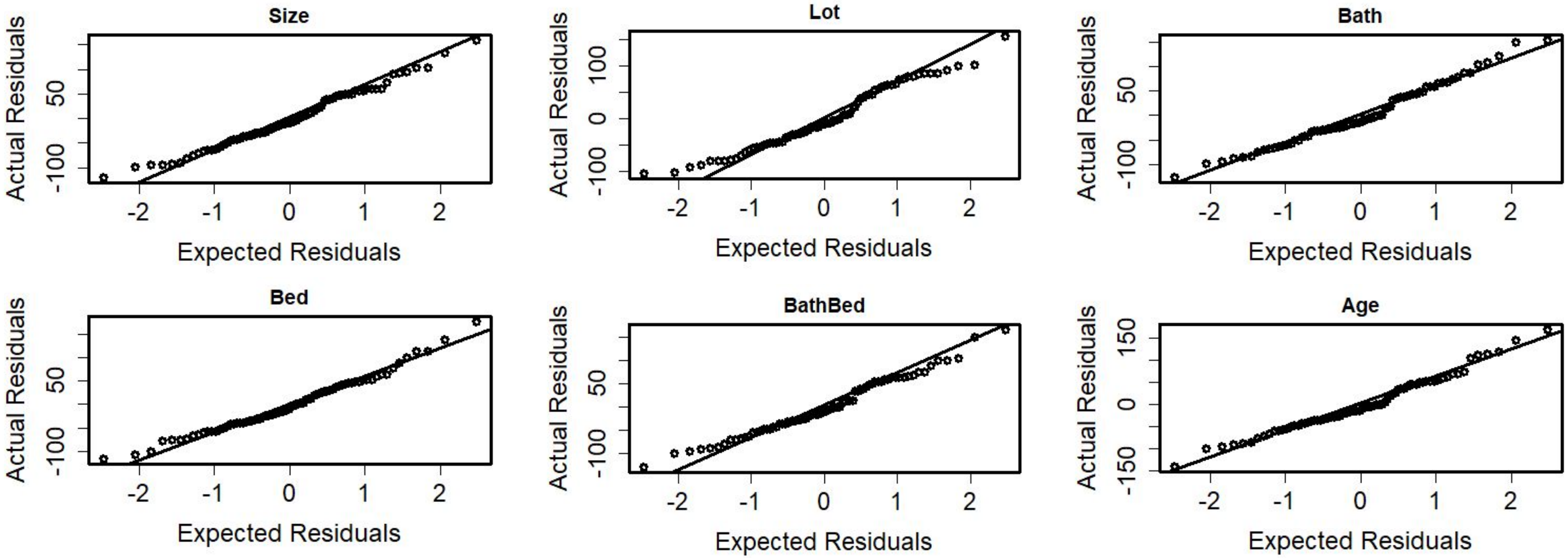




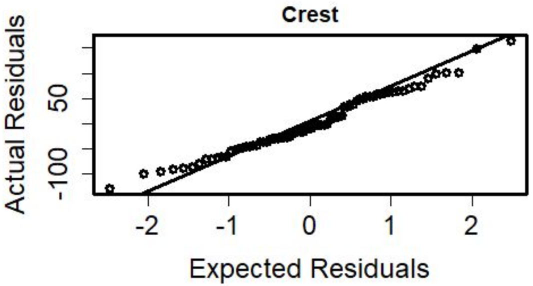


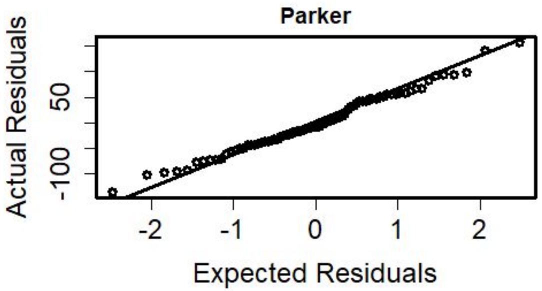


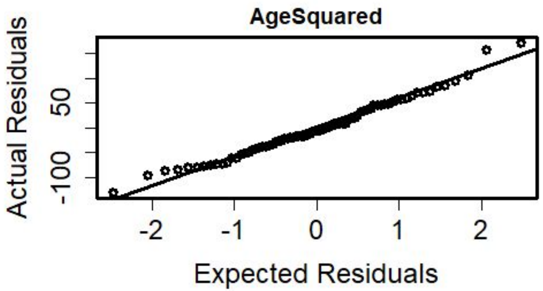
Normal Plots

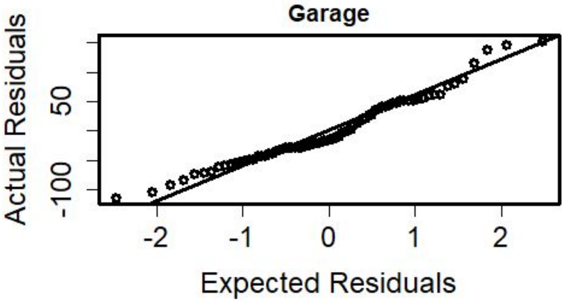


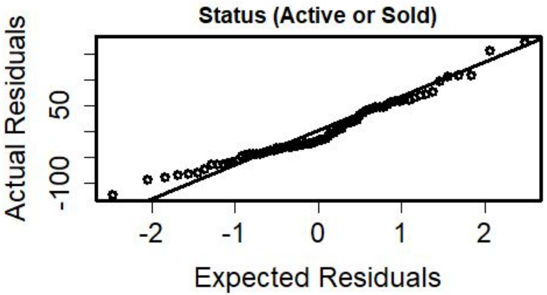
Normal Plots

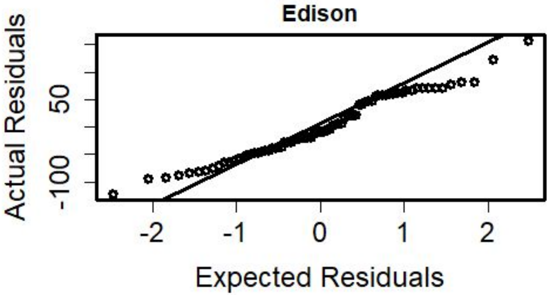


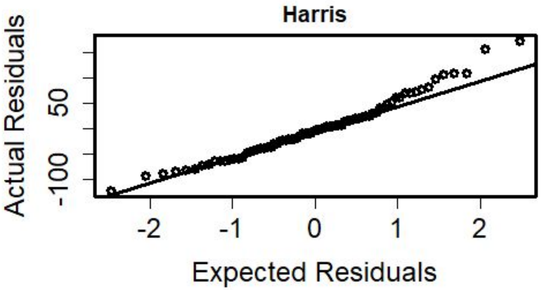


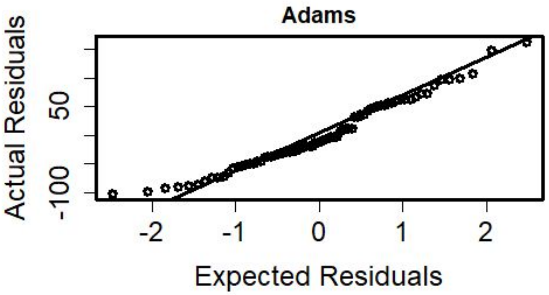






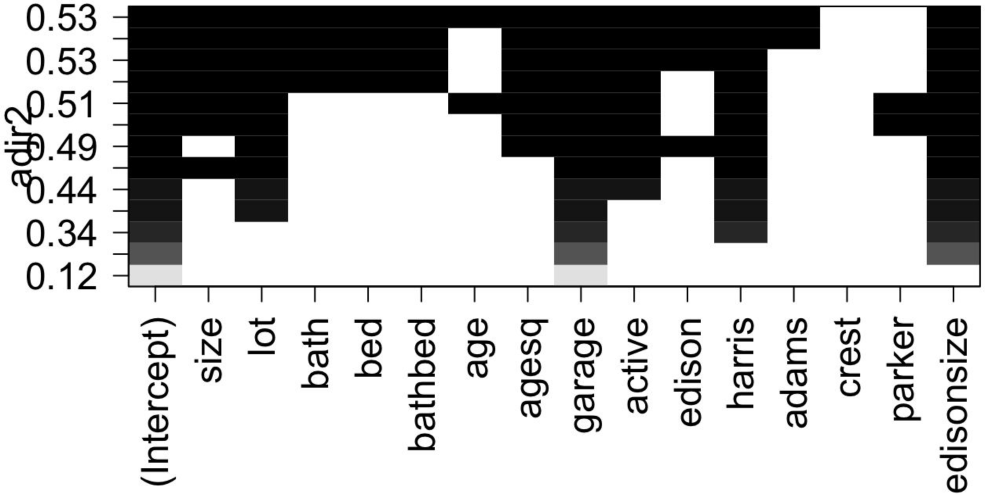






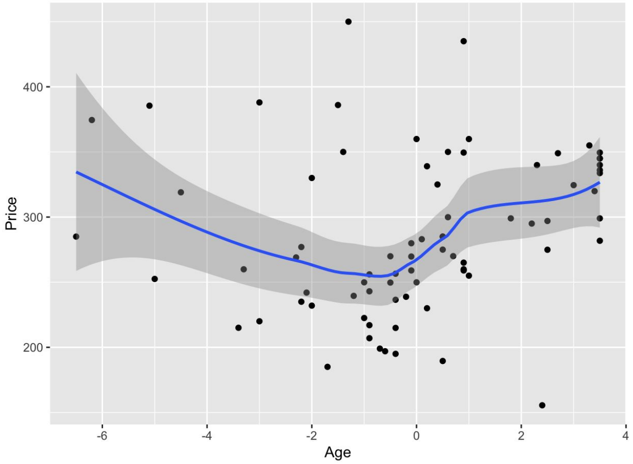
Interaction Terms

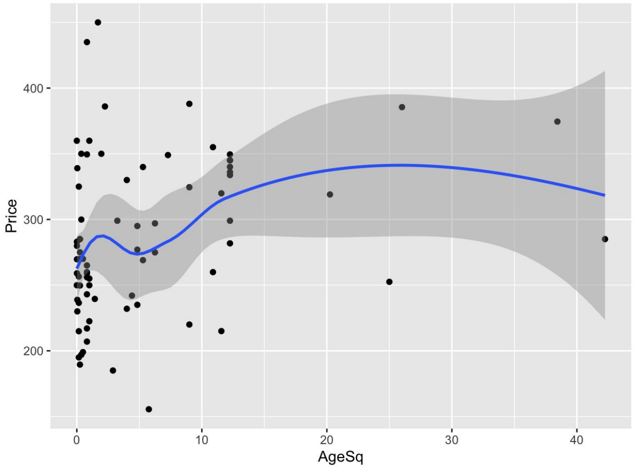
- We included two additional interaction terms
    - bed*bath
    - edison*size



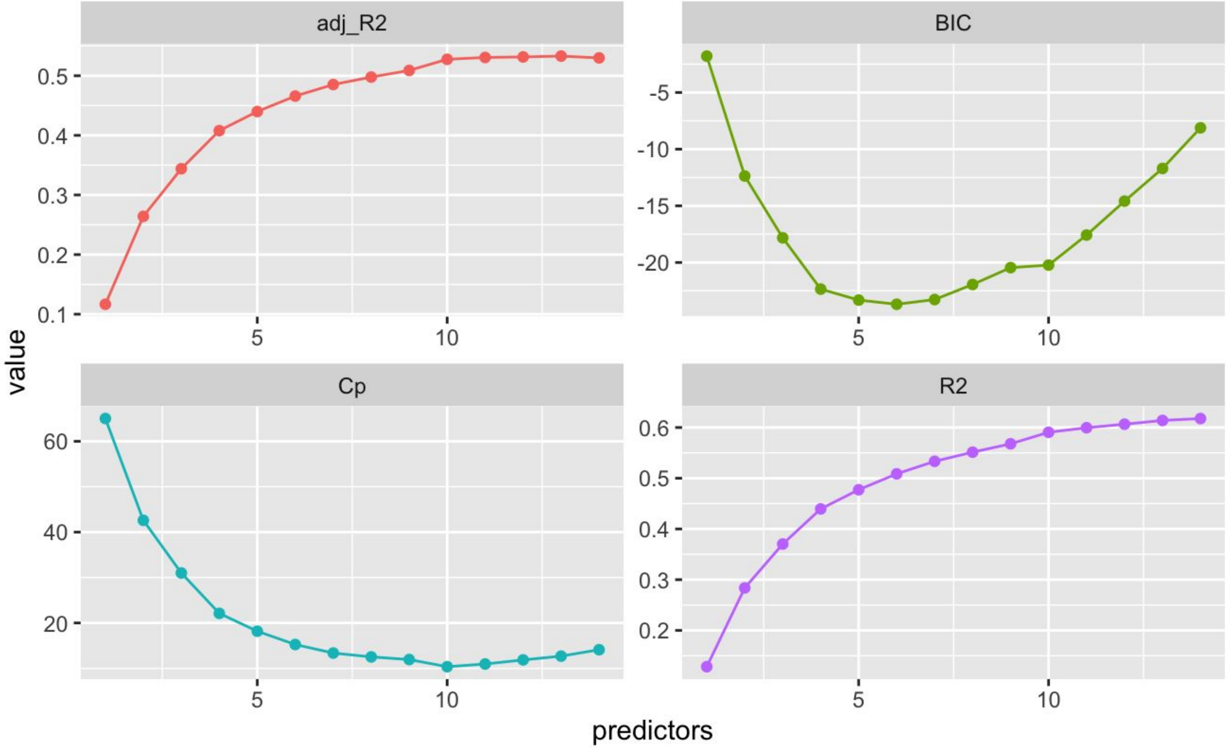
Transformations

- We included a single transform, age^2





Model Selection



The Model

```
price = b0 + b1*size + b2*lot + b3*bath + b4*bed +
        b5*bathbed + b6*agesq + b7*garage +
  b8*active + b9*edison + b10*harris + b11*edisonsize

price = 355.151 + 43.773*size + 8.630*lot - 96.426*bath -
        78.393*bed + 29.347*bathbed + 1.409*agesq +
      20.203*garage + 29.040*active - 149.353*edison +
            50.222*harris + 110.178*edisonsize
```

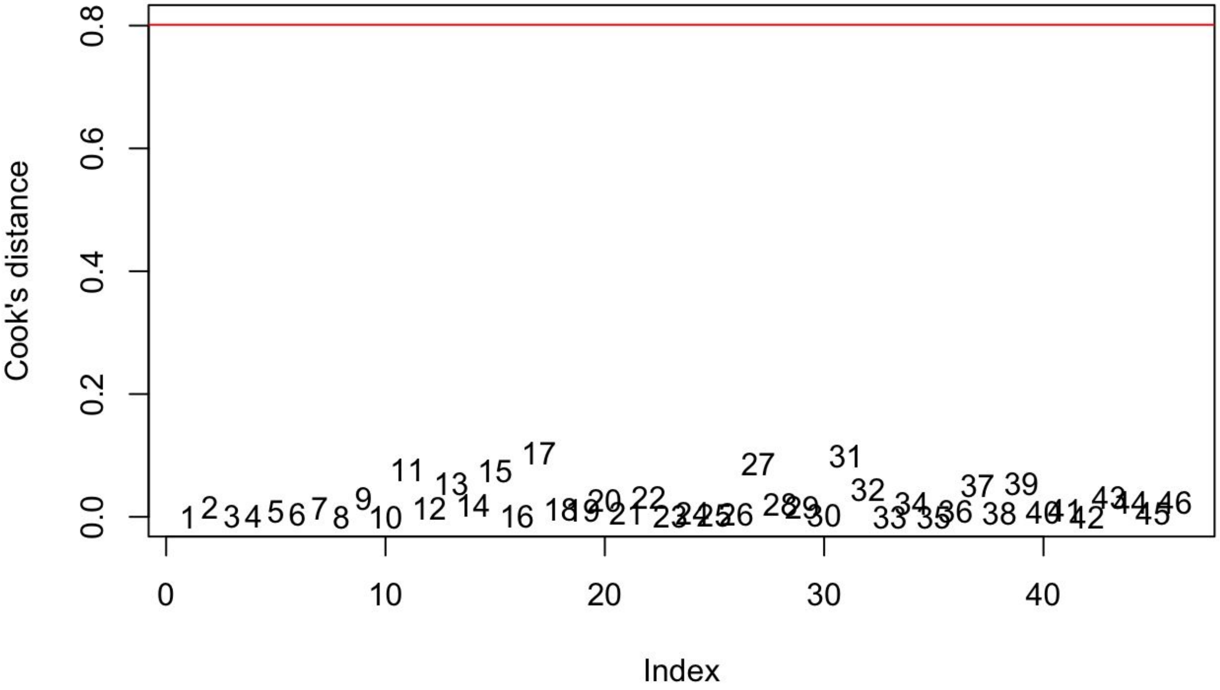
Model Fit

```
Residuals:
    Min      1Q  Median      3Q     Max 
-106.97  -20.91    0.87   23.73   96.69 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  355.151    106.441   3.337 0.001416 ** 
size          43.773     28.981   1.510 0.135861    
lot            8.630      3.364   2.565 0.012656 *  
bath         -96.426     41.831  -2.305 0.024412 *  
bed          -78.393     27.372  -2.864 0.005651 ** 
bathbed       29.347     11.688   2.511 0.014575 *  
agesq          1.409      0.702   2.008 0.048909 *  
garage        20.203      7.973   2.534 0.013741 *  
active        29.040     10.860   2.674 0.009499 ** 
edison      -149.353    124.900  -1.196 0.236192    
harris        50.222     14.418   3.483 0.000898 ***
edisonsize   110.178     64.593   1.706 0.092907 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.33 on 64 degrees of freedom
Multiple R-squared:  0.5995,Adjusted R-squared:  0.5307 
F-statistic:  8.71 on 11 and 64 DF,  p-value: 3.445e-09
```

Cross Validation

- 70/30 train/validation split
- MSPR: 1611.11
- MSE: 1438.56
- → *MSE* for the selected regression model is not seriously biased
- Achieved minimum MSE from all tested models

No Outliers!


Cook's distance
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

Limitations

- We have too little data, there are only 76 data points
- There are only 53 points to fit to with a 70% training split
- Data only representative of Eugene, Oregon
- Data seems to have been pre-processed