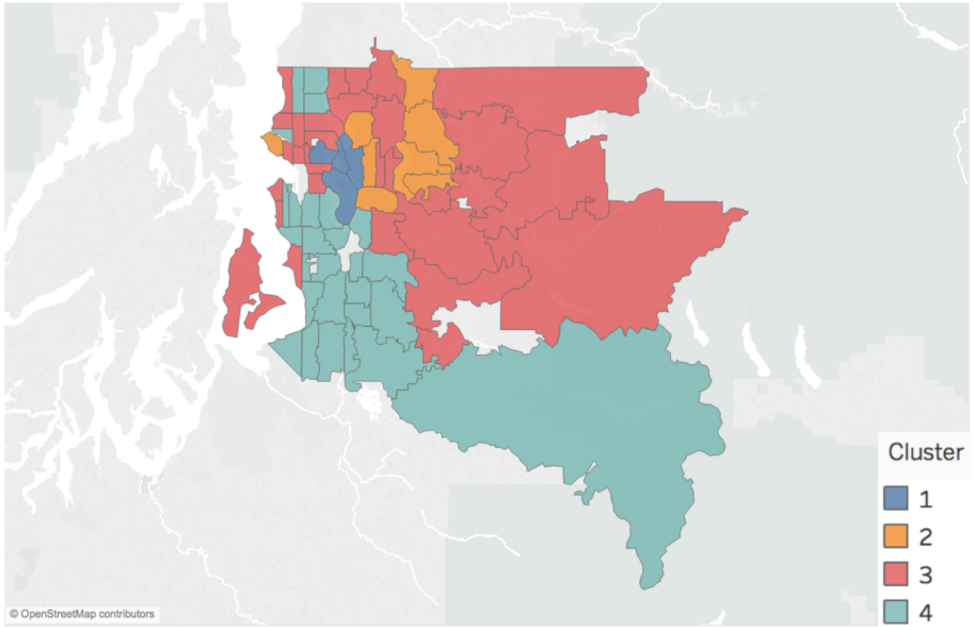
STT805 – Modeling House Prices in King County

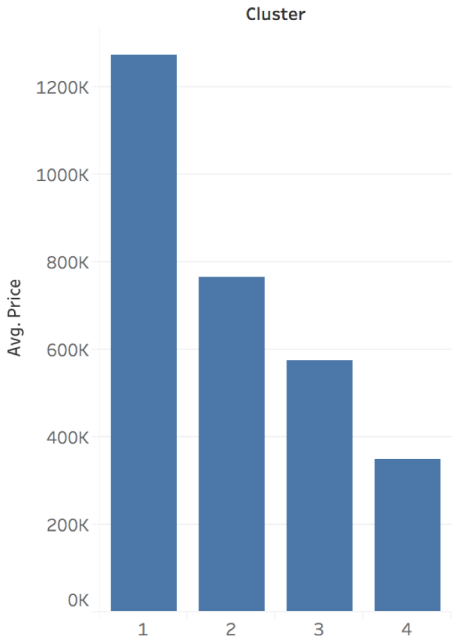
Abhinandan Garg
Shashank Shekhar Katyayan
Derek Herincx
Leonardo de Costa Sousa
Frederik DeRaedt

Project Objective

- Analyze the King County house sales dataset to develop model for understanding and predicting house prices
- Perform linear regression to develop the model and use diagnostic statistical aids to optimize the model
- Develop an information support tool for a house buyer in King County

King Country Map Distribution





Data Overview

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	avg_total_income	age	sqft_basement
221900	3	1.00	1180	5650	1.0	0	0	3	7	1180	51.52818	63	0
538000	3	2.25	2570	7242	2.0	0	0	3	7	2170	65.41865	67	400
180000	2	1.00	770	10000	1.0	0	0	3	6	770	88.18410	85	0
604000	4	3.00	1960	5000	1.0	0	0	5	7	1050	107.07258	53	910
510000	3	2.00	1680	8080	1.0	0	0	3	8	1680	168.39324	31	0
1230000	4	4.50	5420	101930	1.0	0	0	3	11	3890	166.09406	17	1530
257500	3	2.25	1715	6819	2.0	0	0	3	7	1715	48.94220	23	0
291850	3	1.50	1060	9711	1.0	0	0	3	7	1060	53.39661	55	0
229500	3	1.00	1780	7470	1.0	0	0	3	7	1050	58.63936	58	730
323000	3	2.50	1890	6560	2.0	0	0	3	7	1890	88.89898	15	0
662500	3	2.50	3560	9796	1.0	0	0	3	8	1860	80.82738	53	1700
468000	2	1.00	1160	6000	1.0	0	0	4	7	860	102.62093	76	300
310000	3	1.00	1430	19901	1.5	0	0	4	7	1430	88.18410	91	0
400000	3	1.75	1370	9680	1.0	0	0	4	7	1370	168.39324	41	0
530000	5	2.00	1810	4850	1.5	0	0	3	7	1810	78.87528	118	0

Kings County House Sales

Data Overview

- Sample Frame
  - Data found online contains house prices and other house characteristics
  - Income Data was joined with the house prices data
- Dependent Variable
  - Price
- Independent Variables
  - Bedrooms, Bathrooms, Square Footage, Number of Floors, Waterfront, Number of Times Viewed, Condition, Grade, Average Total Income, Age
- Preprocessing
  - Remove all data with Price > $2 million

Analytical Approach

- Identified price as the dependent variable; removed zipcode/location from the dataset
- Joined house prices data with income data to better explain the model
- Performed initial data exploration and variable selection
- Produced two analyses
  - Heuristic: using stepwise model selection
  - Optimal: using variable addition and optimization
- Validated the final model with cross validation
- Data was clean – no missing or null values

SPARTANS WILL.

Initial Data Summary

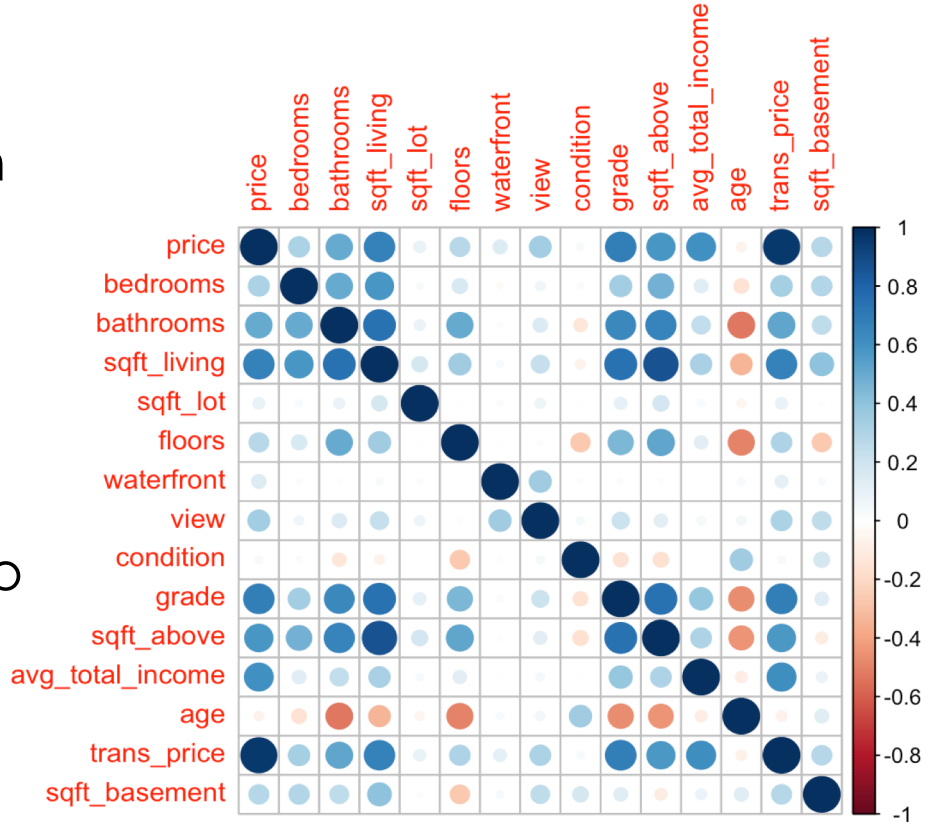
Variables Summary

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition
Min. : 75000	Min. : 0.000	Min. :0.000	Min. : 290	Min. : 520	Min. :1.000	Min. :0.000000	Min. :0.0000	Min. :1.000
1st Qu.: 320000	1st Qu.: 3.000	1st Qu.:1.500	1st Qu.:1420	1st Qu.: 5027	1st Qu.:1.000	1st Qu.:0.000000	1st Qu.:0.0000	1st Qu.:3.000
Median : 450000	Median : 3.000	Median :2.250	Median :1900	Median : 7577	Median :1.500	Median :0.000000	Median :0.0000	Median :3.000
Mean : 518815	Mean : 3.362	Mean :2.098	Mean :2051	Mean : 14993	Mean :1.491	Mean :0.005326	Mean :0.2167	Mean :3.408
3rd Qu.: 635000	3rd Qu.: 4.000	3rd Qu.:2.500	3rd Qu.:2520	3rd Qu.: 10550	3rd Qu.:2.000	3rd Qu.:0.000000	3rd Qu.:0.0000	3rd Qu.:4.000
Max. :1990000	Max. :33.000	Max. :7.500	Max. :7730	Max. :1651359	Max. :3.500	Max. :1.000000	Max. :4.0000	Max. :5.000

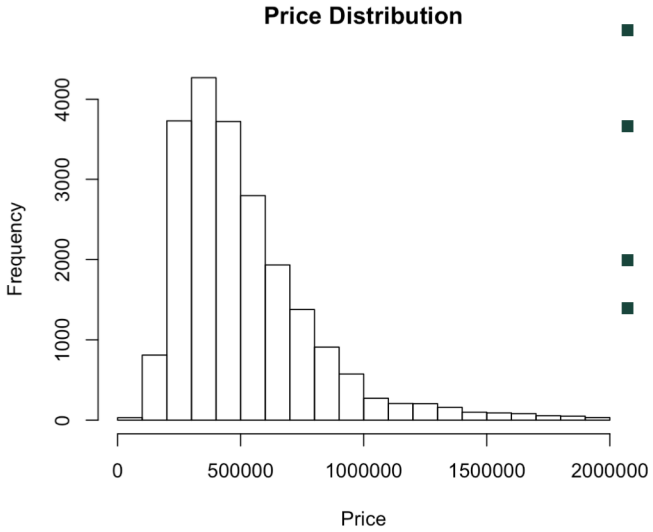
grade	sqft_above	avg_total_income	age	trans_price	sqft_basement
Min. : 1.000	Min. : 290	Min. : 40.60	Min. : 3.00	Min. : 9.441	Min. : 0.0
1st Qu.: 7.000	1st Qu.:1190	1st Qu.: 65.42	1st Qu.: 21.00	1st Qu.:12.619	1st Qu.: 0.0
Median : 7.000	Median :1550	Median : 86.91	Median : 43.00	Median :13.510	Median : 0.0
Mean : 7.627	Mean :1767	Mean : 96.61	Mean : 47.01	Mean :13.616	Mean : 284.5
3rd Qu.: 8.000	3rd Qu.:2190	3rd Qu.:113.65	3rd Qu.: 67.00	3rd Qu.:14.473	3rd Qu.: 550.0
Max. :13.000	Max. :7420	Max. :547.75	Max. :118.00	Max. :18.187	Max. :3260.0

Correlation Matrix

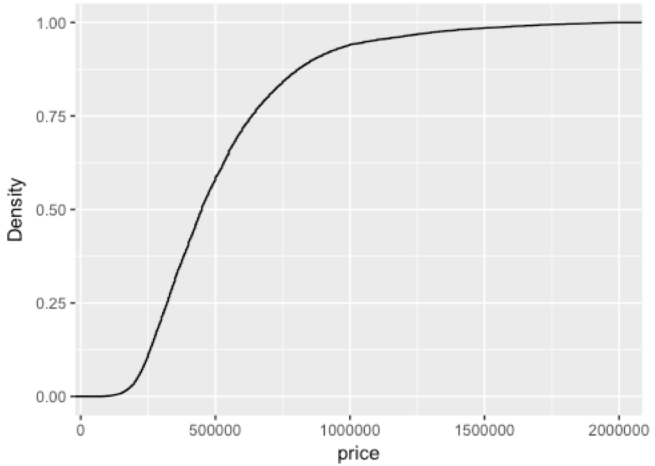
- Heavy correlation between some variables specially in terms of space: such as sqft_lot with bedrooms and bathrooms
- Price is heavily correlated to several variables



Price Distribution



- Price distribution is skewed right
- Only a very small percentage of data is greater than $1million
- Mean: $518,000
- Median: $450,000



SPARTANS WILL.

Model Development

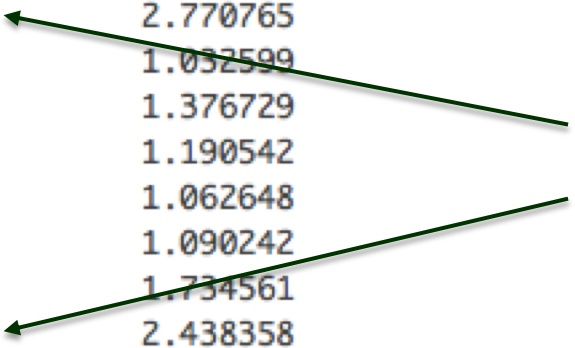
Linear model with all the variables

```
Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.515e+05  1.154e+04 -73.802  < 2e-16 ***
bedrooms        -1.526e+04  1.346e+03 -11.339  < 2e-16 ***
bathrooms        3.151e+04  2.302e+03  13.687  < 2e-16 ***
sqft_living      1.164e+02  3.110e+00  37.420  < 2e-16 ***
sqft_lot        -1.891e-02  2.440e-02  -0.775    0.438
floors           4.924e+04  2.472e+03  19.923  < 2e-16 ***
waterfront1      2.225e+05  1.584e+04  14.051  < 2e-16 ***
view1            8.055e+04  8.067e+03   9.985  < 2e-16 ***
view2            7.393e+04  4.867e+03  15.190  < 2e-16 ***
view3            1.422e+05  6.747e+03  21.079  < 2e-16 ***
view4            2.478e+05  1.085e+04  22.841  < 2e-16 ***
condition        1.255e+04  1.625e+03   7.728 1.14e-14 ***
grade            8.787e+04  1.477e+03  59.474  < 2e-16 ***
sqft_above      -2.124e+01  2.979e+00  -7.130 1.04e-12 ***
avg_total_income 2.326e+03  2.311e+01 100.625  < 2e-16 ***
age              2.804e+03  4.478e+01  62.607  < 2e-16 ***
sqft_basement           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141600 on 21389 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7553
F-statistic:  4406 on 15 and 21389 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.515e+05  1.154e+04 -73.802  < 2e-16 ***
bedrooms        -1.526e+04  1.346e+03 -11.339  < 2e-16 ***
bathrooms        3.151e+04  2.302e+03  13.687  < 2e-16 ***
sqft_living      1.164e+02  3.110e+00  37.420  < 2e-16 ***
sqft_lot        -1.891e-02  2.440e-02  -0.775    0.438
floors           4.924e+04  2.472e+03  19.923  < 2e-16 ***
waterfront1      2.225e+05  1.584e+04  14.051  < 2e-16 ***
view1            8.055e+04  8.067e+03   9.985  < 2e-16 ***
view2            7.393e+04  4.867e+03  15.190  < 2e-16 ***
view3            1.422e+05  6.747e+03  21.079  < 2e-16 ***
view4            2.478e+05  1.085e+04  22.841  < 2e-16 ***
condition        1.255e+04  1.625e+03   7.728 1.14e-14 ***
grade            8.787e+04  1.477e+03  59.474  < 2e-16 ***
sqft_above      -2.124e+01  2.979e+00  -7.130 1.04e-12 ***
avg_total_income 2.326e+03  2.311e+01 100.625  < 2e-16 ***
age              2.804e+03  4.478e+01  62.607  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141600 on 21389 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7553
F-statistic:  4406 on 15 and 21389 DF,  p-value: < 2.2e-16
```

VIF (model with all variables)

	GVIF	Df	GVIF^(1/(2*Df))
bedrooms	1.651926	1	1.285273
bathrooms	3.157880	1	1.777042
sqft_living	7.677140	1	2.770765
sqft_lot	1.066260	1	1.032599
floors	1.895383	1	1.376729
waterfront	1.417390	1	1.190542
view	1.625982	4	1.062648
condition	1.188629	1	1.090242
grade	3.008703	1	1.734561
sqft_above	5.945591	1	2.438358
avg_total_income	1.202087	1	1.096397
age	1.837997	1	1.355728



Signs of multicollinearity

Model Selection and Validation

```
Selection Algorithm: exhaustive
          bedrooms bathrooms waterfront view floors grade sqft_lot age avg_total_income condition sqft_above
1  ( 1 )  " "      " "       " "        " "  " "    "*"   " "      " " " "                " "       " "
2  ( 1 )  " "      " "       " "        " "  " "    "*"   " "      " " "*"                " "       " "
3  ( 1 )  " "      " "       " "        " "  " "    "*"   " "      "*" "*"                " "       " "
4  ( 1 )  " "      "*"       " "        " "  " "    "*"   " "      "*" "*"                " "       " "
5  ( 1 )  " "      "*"       " "        "*"  " "    "*"   " "      "*" "*"                " "       " "
6  ( 1 )  " "      "*"       " "        "*"  " "    "*"   " "      "*" "*"                " "       "*"
7  ( 1 )  " "      "*"       "*"        "*"  " "    "*"   " "      "*" "*"                " "       "*"
8  ( 1 )  " "      "*"       "*"        "*"  " "    "*"   " "      "*" "*"                "*"       "*"
9  ( 1 )  " "      "*"       "*"        "*"  "*"    "*"   " "      "*" "*"                "*"       "*"
10  ( 1 ) "*"      "*"       "*"        "*"  "*"    "*"   " "      "*" "*"                "*"       "*"
11  ( 1 ) "*"      "*"       "*"        "*"  "*"    "*"   "*"      "*" "*"                "*"       "*"
```

Model Selection and Validation

	Rp2	Rap2	C_p	SSE_p	BIC
1	0.4630328	0.4630077	22527.43242	9.423643e+14	-13290.08
2	0.6032080	0.6031709	11061.91780	6.963601e+14	-19755.65
3	0.6641740	0.6641269	6076.39260	5.893664e+14	-23316.44
4	0.6970675	0.6970109	3387.42149	5.316390e+14	-25512.97
5	0.7229811	0.7229164	1269.47110	4.861613e+14	-27417.12
6	0.7333250	0.7332502	425.25572	4.680081e+14	-28221.72
7	0.7368597	0.7367736	138.09184	4.618048e+14	-28497.35
8	0.7376803	0.7375822	72.95389	4.603646e+14	-28554.24
9	0.7384879	0.7383779	8.88585	4.589473e+14	-28610.27
10	0.7384969	0.7383747	10.15125	4.589315e+14	-28601.04
11	0.7384987	0.7383643	12.00000	4.589283e+14	-28591.22

Regression Summary

- Majority of the variables provided in the dataset were significant
- The % standard error varies on parameters differently

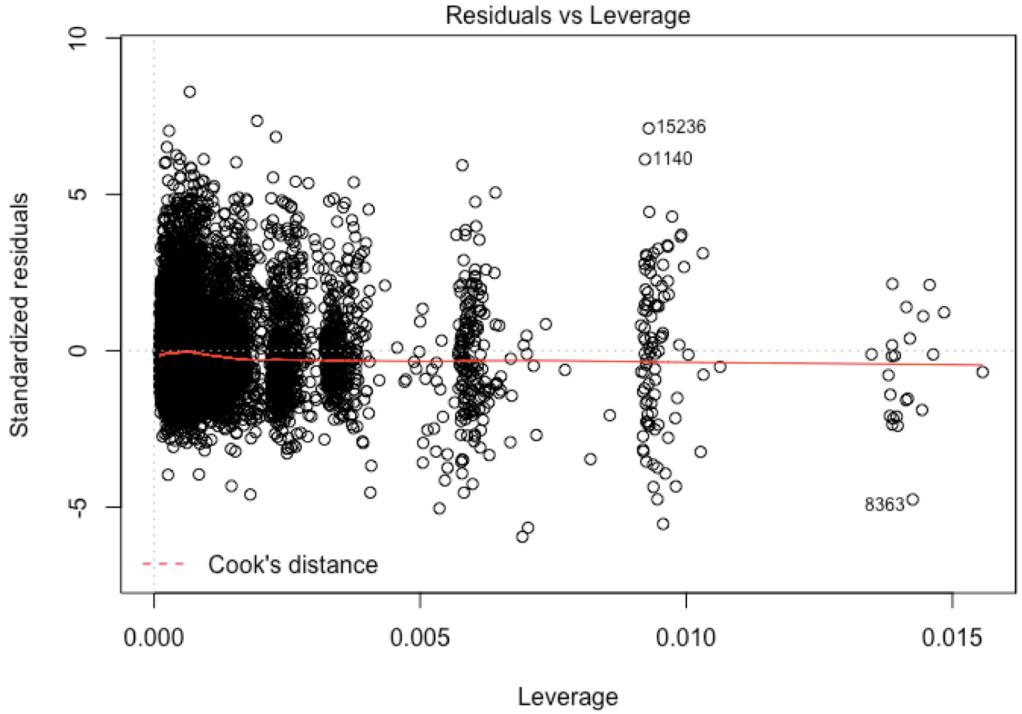
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.701e+05  1.116e+04 -86.931   <2e-16 ***
bathrooms        6.680e+04  2.014e+03  33.174   <2e-16 ***
floors           1.962e+04  2.373e+03   8.268   <2e-16 ***
waterfront1      2.052e+05  1.633e+04  12.568   <2e-16 ***
view1            1.037e+05  8.301e+03  12.487   <2e-16 ***
view2            9.434e+04  4.989e+03  18.909   <2e-16 ***
view3            1.752e+05  6.891e+03  25.422   <2e-16 ***
view4            2.914e+05  1.114e+04  26.169   <2e-16 ***
condition        1.529e+04  1.671e+03   9.149   <2e-16 ***
grade            1.025e+05  1.467e+03  69.906   <2e-16 ***
sqft_above       5.607e+01  2.086e+00  26.881   <2e-16 ***
avg_total_income 2.353e+03  2.382e+01  98.790   <2e-16 ***
age              3.077e+03  4.528e+01  67.959   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 146200 on 21392 degrees of freedom
Multiple R-squared:  0.7395,    Adjusted R-squared:  0.7393
F-statistic:  5060 on 12 and 21392 DF,  p-value: < 2.2e-16
```

Multicollinearity using VIF

	GVIF	Df	GVIF^(1/(2*Df))
bathrooms	2.267998	1	1.505987
floors	1.640449	1	1.280800
waterfront	1.413927	1	1.189087
view	1.554336	4	1.056679
condition	1.181055	1	1.086764
grade	2.784011	1	1.668536
sqft_above	2.735774	1	1.654017
avg_total_income	1.198681	1	1.094843
age	1.764033	1	1.328169

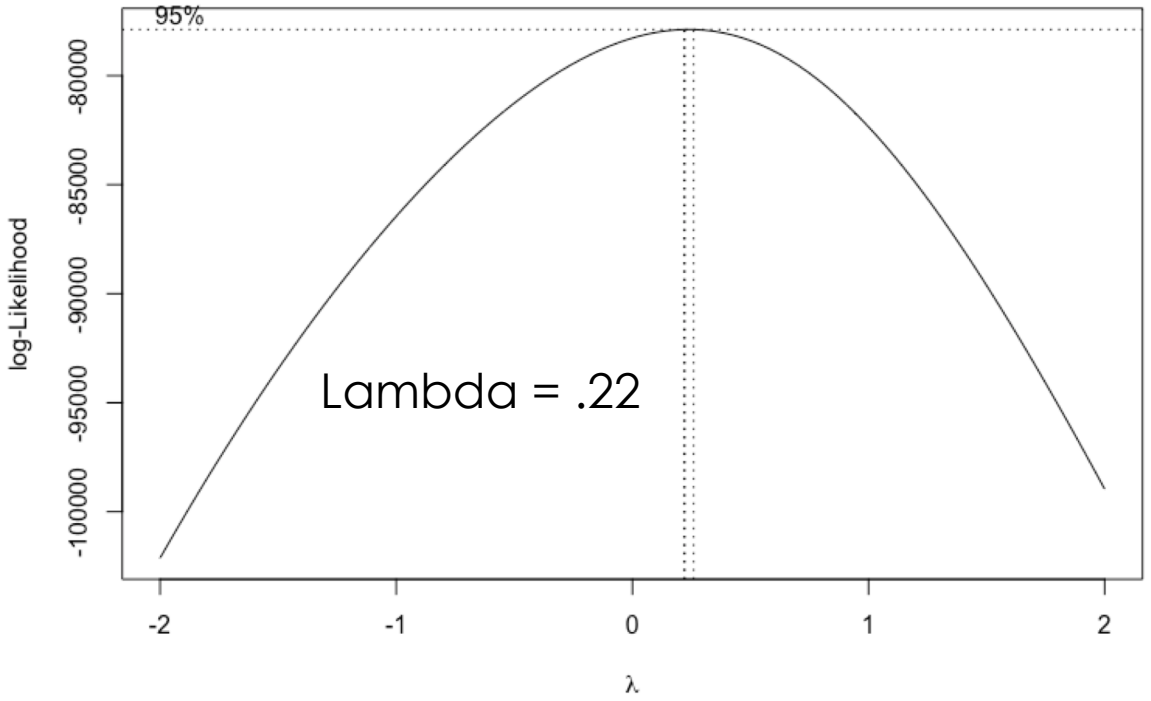
Multicollinearity has been accounted for

Residuals vs Leverage



No highly influential observations

Box-Cox Normality Plot



Transformation of dependent variable

$$Price' = Price^{0.2}$$

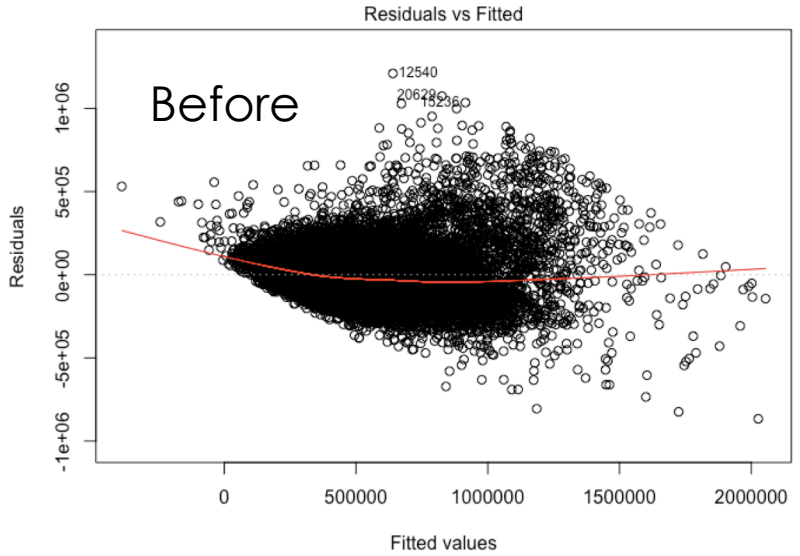
➔ $\text{Price}' = 6.17 + 0.368 * \text{bathrooms} + 0.177 * \text{floors} + 0.763*\text{waterfront1} + 0.475*\text{view1} + 0.434*\text{view2} + 0.638*\text{view3} + 1.408*\text{view4} + 0.090*\text{condition} + 0.513*\text{grade} + 1.813*10^{-4} * \text{sqft\_avg} + 1.16 * 10^{-2} * \text{avg\_total\_income} + 1.47 * 10^{-2} * \text{age}$

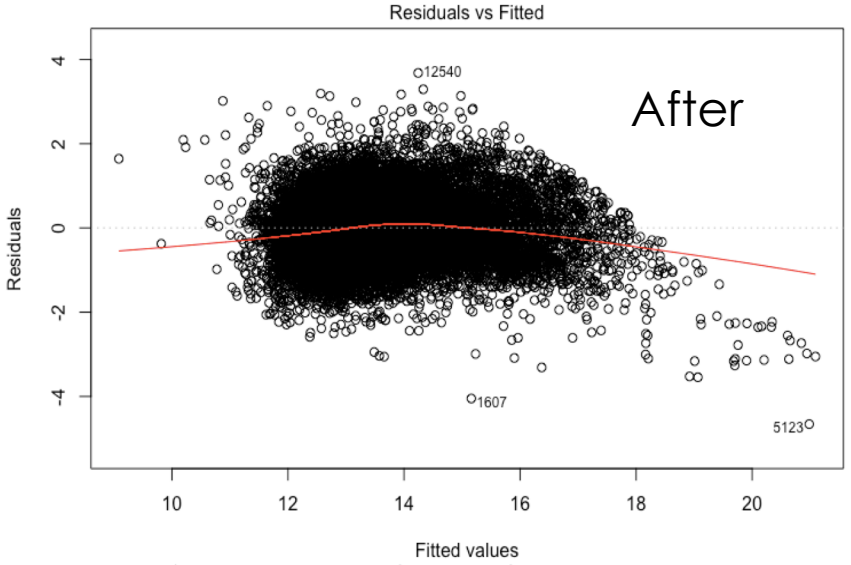
Regression Summary II

- New model is performed with a transformed price variable
- The variance explained is still the same while diagnostics show improvement in explanatory ability

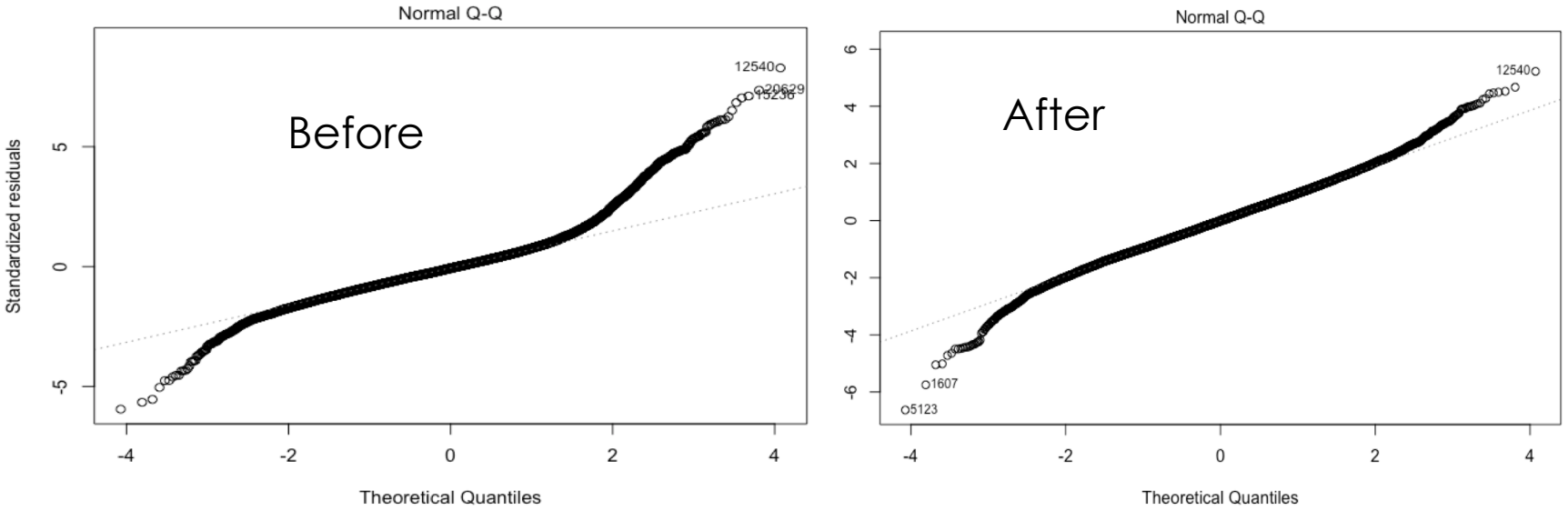
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.170e+00  5.381e-02 114.669   <2e-16 ***
bathrooms        3.676e-01  9.709e-03  37.865   <2e-16 ***
floors           1.774e-01  1.144e-02  15.499   <2e-16 ***
waterfront1      7.630e-01  7.872e-02   9.693   <2e-16 ***
view1            4.754e-01  4.002e-02  11.878   <2e-16 ***
view2            4.344e-01  2.406e-02  18.057   <2e-16 ***
view3            6.379e-01  3.322e-02  19.200   <2e-16 ***
view4            1.048e+00  5.369e-02  19.528   <2e-16 ***
condition        8.962e-02  8.059e-03  11.120   <2e-16 ***
grade            5.132e-01  7.073e-03  72.561   <2e-16 ***
sqft_above       1.813e-04  1.006e-05  18.025   <2e-16 ***
avg_total_income 1.160e-02  1.149e-04 100.956   <2e-16 ***
age              1.474e-02  2.183e-04  67.498   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7049 on 21392 degrees of freedom
Multiple R-squared:  0.7378,    Adjusted R-squared:  0.7377
F-statistic:  5017 on 12 and 21392 DF,  p-value: < 2.2e-16
```

Residuals vs Fitted Values





QQ plot Comparison



Residual distribution is better with the new model

SPARTANS WILL.

Heuristic Approach

Heuristic Approach - Stepwise with Price

```
Call:
lm(formula = price ~ grade + avg_total_income + age + sqft_living +
    view + floors + waterfront + bathrooms + bedrooms + condition +
    sqft_above, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-882808  -81613  -10120   66015 1172898

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.521e+05  1.151e+04 -74.059  < 2e-16 ***
grade           8.791e+04  1.476e+03  59.552  < 2e-16 ***
avg_total_income 2.326e+03 2.311e+01 100.673  < 2e-16 ***
age             2.804e+03  4.477e+01  62.646  < 2e-16 ***
sqft_living     1.163e+02  3.108e+00  37.418  < 2e-16 ***
view1           8.063e+04  8.066e+03   9.996  < 2e-16 ***
view2           7.384e+04  4.865e+03  15.176  < 2e-16 ***
view3           1.419e+05  6.736e+03  21.068  < 2e-16 ***
view4           2.479e+05  1.085e+04  22.845  < 2e-16 ***
floors          4.944e+04  2.458e+03  20.116  < 2e-16 ***
waterfront1     2.224e+05  1.583e+04  14.043  < 2e-16 ***
bathrooms       3.153e+04  2.302e+03  13.699  < 2e-16 ***
bedrooms       -1.517e+04  1.341e+03 -11.313  < 2e-16 ***
condition       1.255e+04  1.625e+03   7.723 1.19e-14 ***
sqft_above     -2.150e+01  2.960e+00  -7.266 3.83e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141600 on 21390 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7553
F-statistic:  4721 on 14 and 21390 DF,  p-value: < 2.2e-16
```

Forward

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + floors +
    waterfront + view + condition + grade + sqft_above + avg_total_income +
    age, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-882808  -81613  -10120   66015 1172898

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.521e+05  1.151e+04 -74.059  < 2e-16 ***
bedrooms       -1.517e+04  1.341e+03 -11.313  < 2e-16 ***
bathrooms       3.153e+04  2.302e+03  13.699  < 2e-16 ***
sqft_living     1.163e+02  3.108e+00  37.418  < 2e-16 ***
floors          4.944e+04  2.458e+03  20.116  < 2e-16 ***
waterfront1     2.224e+05  1.583e+04  14.043  < 2e-16 ***
view1           8.063e+04  8.066e+03   9.996  < 2e-16 ***
view2           7.384e+04  4.865e+03  15.176  < 2e-16 ***
view3           1.419e+05  6.736e+03  21.068  < 2e-16 ***
view4           2.479e+05  1.085e+04  22.845  < 2e-16 ***
condition       1.255e+04  1.625e+03   7.723 1.19e-14 ***
grade           8.791e+04  1.476e+03  59.552  < 2e-16 ***
sqft_above     -2.150e+01  2.960e+00  -7.266 3.83e-13 ***
avg_total_income 2.326e+03 2.311e+01 100.673  < 2e-16 ***
age             2.804e+03  4.477e+01  62.646  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141600 on 21390 degrees of freedom
Multiple R-squared:  0.7555,    Adjusted R-squared:  0.7553
F-statistic:  4721 on 14 and 21390 DF,  p-value: < 2.2e-16
```

Backward

Heuristic Approach - Stepwise with Transformed Price

```
Call:
lm(formula = trans_price ~ grade + avg_total_income + age + sqft_living +
    view + floors + bathrooms + sqft_above + waterfront + condition +
    bedrooms + sqft_lot, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2864 -0.4383  0.0000  0.4470  3.5130

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.665e+00  5.562e-02 119.819  < 2e-16 ***
grade           4.460e-01  7.123e-03  62.620  < 2e-16 ***
avg_total_income 1.149e-02  1.114e-04 103.140  < 2e-16 ***
age             1.332e-02  2.159e-04  61.691  < 2e-16 ***
sqft_living     5.561e-04  1.499e-05  37.089  < 2e-16 ***
view1           3.680e-01  3.889e-02   9.463  < 2e-16 ***
view2           3.376e-01  2.346e-02  14.388  < 2e-16 ***
view3           4.787e-01  3.253e-02  14.719  < 2e-16 ***
view4           8.432e-01  5.231e-02  16.119  < 2e-16 ***
floors          3.314e-01  1.192e-02  27.809  < 2e-16 ***
bathrooms       1.801e-01  1.110e-02  16.227  < 2e-16 ***
sqft_above     -2.085e-04  1.436e-05 -14.517  < 2e-16 ***
waterfront1     8.580e-01  7.635e-02  11.238  < 2e-16 ***
condition       7.373e-02  7.832e-03   9.413  < 2e-16 ***
bedrooms       -3.936e-02  6.488e-03  -6.066 1.33e-09 ***
sqft_lot        2.732e-07  1.176e-07   2.323   0.0202 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6829 on 21389 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.7538
F-statistic:  4371 on 15 and 21389 DF,  p-value: < 2.2e-16
```

Forward

```
Call:
lm(formula = trans_price ~ bedrooms + bathrooms + sqft_living +
    sqft_lot + floors + waterfront + view + condition + grade +
    sqft_above + avg_total_income + age, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2864 -0.4383  0.0000  0.4470  3.5130

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.665e+00  5.562e-02 119.819  < 2e-16 ***
bedrooms       -3.936e-02  6.488e-03  -6.066 1.33e-09 ***
bathrooms       1.801e-01  1.110e-02  16.227  < 2e-16 ***
sqft_living     5.561e-04  1.499e-05  37.089  < 2e-16 ***
sqft_lot        2.732e-07  1.176e-07   2.323   0.0202 *
floors          3.314e-01  1.192e-02  27.809  < 2e-16 ***
waterfront1     8.580e-01  7.635e-02  11.238  < 2e-16 ***
view1           3.680e-01  3.889e-02   9.463  < 2e-16 ***
view2           3.376e-01  2.346e-02  14.388  < 2e-16 ***
view3           4.787e-01  3.253e-02  14.719  < 2e-16 ***
view4           8.432e-01  5.231e-02  16.119  < 2e-16 ***
condition       7.373e-02  7.832e-03   9.413  < 2e-16 ***
grade           4.460e-01  7.123e-03  62.620  < 2e-16 ***
sqft_above     -2.085e-04  1.436e-05 -14.517  < 2e-16 ***
avg_total_income 1.149e-02  1.114e-04 103.140  < 2e-16 ***
age             1.332e-02  2.159e-04  61.691  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6829 on 21389 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.7538
F-statistic:  4371 on 15 and 21389 DF,  p-value: < 2.2e-16
```

Backward

SPARTANS WILL.

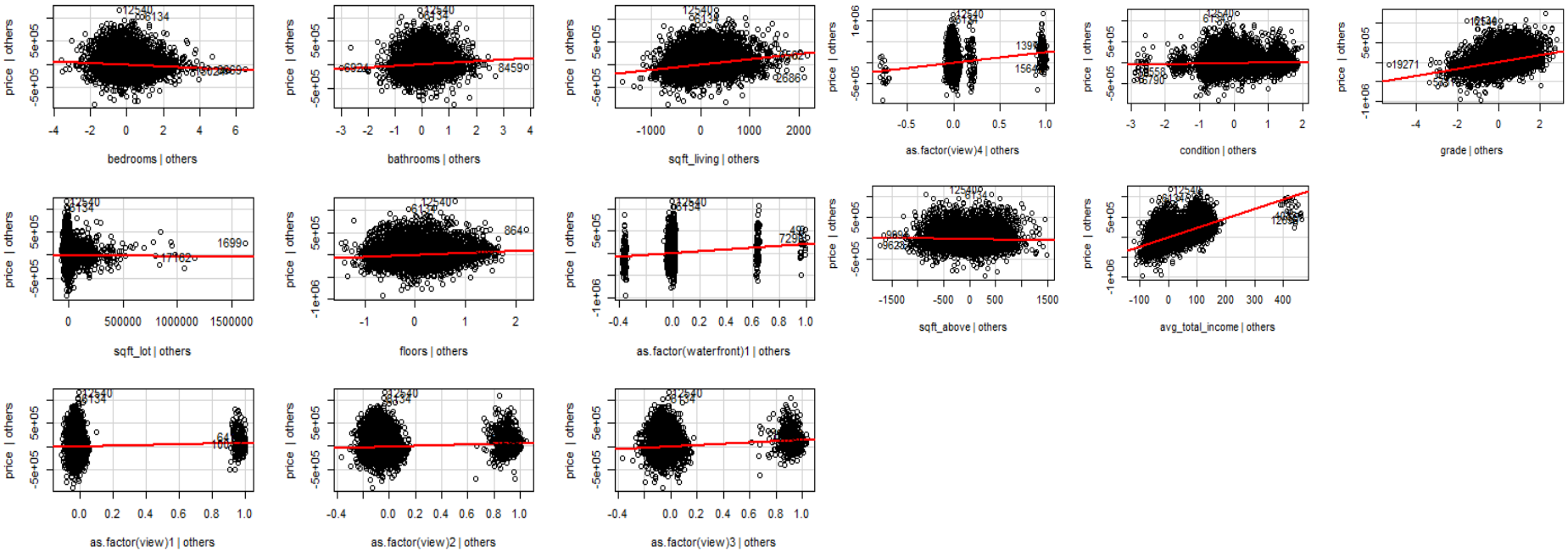
Optimal Model Approach

Variable Addition Method - Optimal

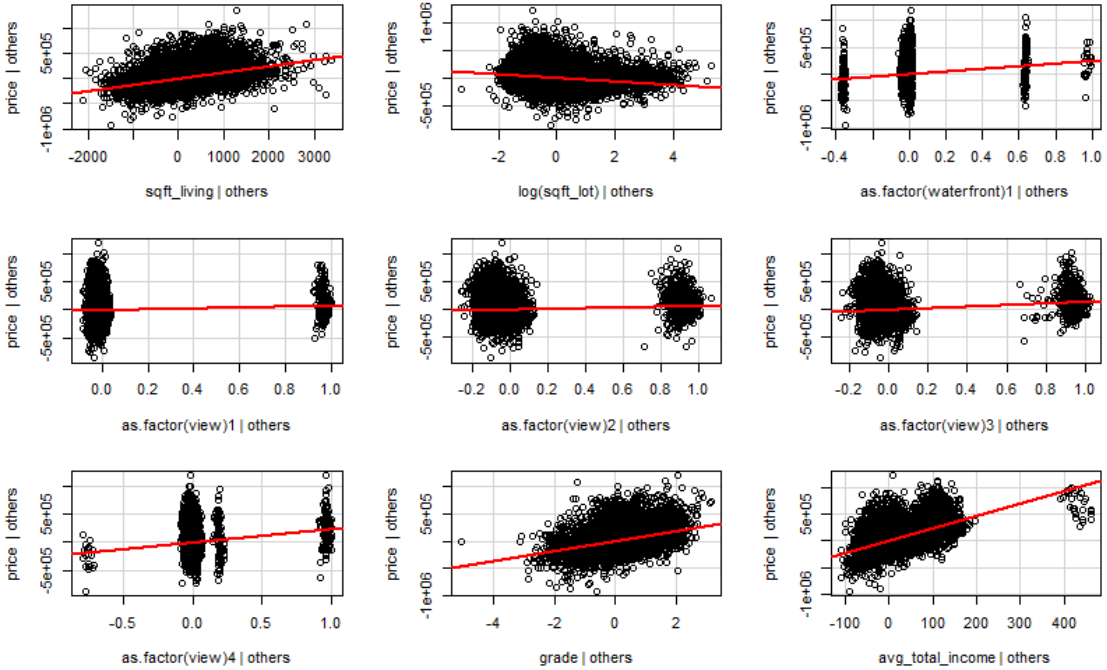
```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + as.factor(waterfront) + as.factor(view) + condition +
    grade + sqft_above + avg_total_income + age, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-886622  -81703   -9883   66095 1171710

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -8.465e+05  1.158e+04 -73.111  < 2e-16 ***
bedrooms            -1.714e+04  1.402e+03 -12.225  < 2e-16 ***
bathrooms            3.213e+04  2.304e+03  13.944  < 2e-16 ***
sqft_living          1.176e+02  3.119e+00  37.701  < 2e-16 ***
sqft_lot            -2.162e-02  2.439e-02  -0.886    0.375
floors               4.910e+04  2.471e+03  19.873  < 2e-16 ***
as.factor(waterfront)1  2.219e+05  1.583e+04  14.020  < 2e-16 ***
as.factor(view)1     8.025e+04  8.063e+03   9.953  < 2e-16 ***
as.factor(view)2     7.363e+04  4.865e+03  15.135  < 2e-16 ***
as.factor(view)3     1.417e+05  6.744e+03  21.015  < 2e-16 ***
as.factor(view)4     2.473e+05  1.085e+04  22.802  < 2e-16 ***
condition            1.254e+04  1.624e+03   7.725 1.17e-14 ***
grade                8.757e+04  1.478e+03  59.254  < 2e-16 ***
sqft_above          -2.121e+01  2.978e+00  -7.122 1.10e-12 ***
avg_total_income     2.324e+03  2.310e+01 100.610  < 2e-16 ***
age                  2.807e+03  4.477e+01  62.714  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141600 on 21388 degrees of freedom
Multiple R-squared:  0.7557,    Adjusted R-squared:  0.7556
F-statistic:  4412 on 15 and 21388 DF,  p-value: < 2.2e-16
```

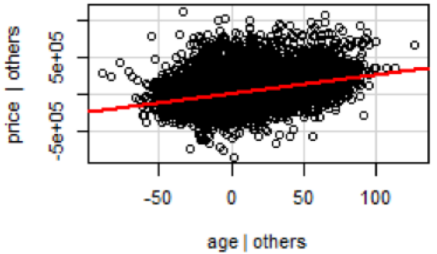
```
Call:
lm(formula = price ~ sqft_living + log(sqft_lot) + as.factor(waterfront) +
    as.factor(view) + grade + avg_total_income + age, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-861248  -81737  -10400   65837 1197515

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.887e+05  1.383e+04  -35.34   <2e-16 ***
sqft_living          1.246e+02  1.811e+00   68.82   <2e-16 ***
log(sqft_lot)       -3.354e+04  1.163e+03  -28.84   <2e-16 ***
as.factor(waterfront)1  2.546e+05  1.586e+04   16.05   <2e-16 ***
as.factor(view)1     8.086e+04  8.057e+03   10.04   <2e-16 ***
as.factor(view)2     7.913e+04  4.847e+03   16.32   <2e-16 ***
as.factor(view)3     1.521e+05  6.697e+03   22.70   <2e-16 ***
as.factor(view)4     2.519e+05  1.082e+04   23.29   <2e-16 ***
grade                9.127e+04  1.426e+03   63.99   <2e-16 ***
avg_total_income     2.343e+03  2.311e+01  101.38   <2e-16 ***
age                  2.515e+03  3.839e+01   65.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141900 on 21393 degrees of freedom
Multiple R-squared:  0.7544,    Adjusted R-squared:  0.7543
F-statistic:  6572 on 10 and 21393 DF,  p-value: < 2.2e-16
```

Optimal Model – Old Model



Optimal Model – New Model



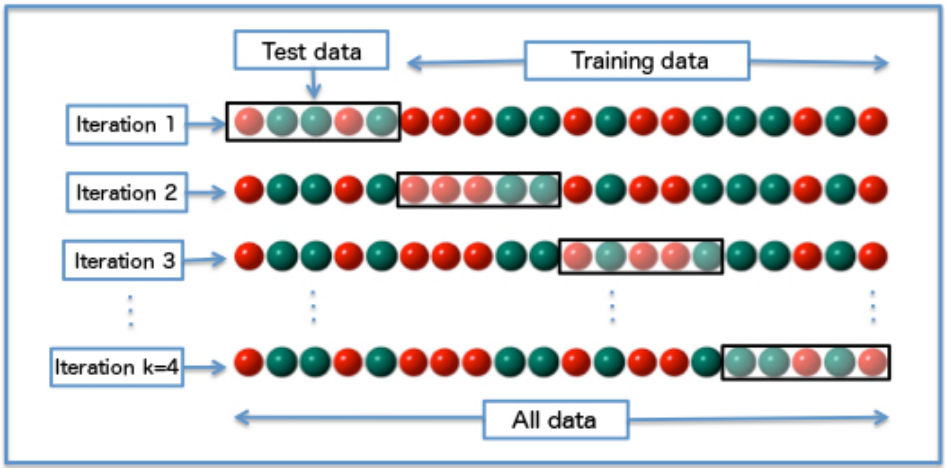


	GVIF	Df	GVIF^(1/(2*Df))
sqft_living	2.592269	1	1.610052
log(sqft_lot)	1.167033	1	1.080293
as.factor(waterfront)	1.415895	1	1.189914
as.factor(view)	1.560521	4	1.057204
grade	2.792675	1	1.671130
avg_total_income	1.196570	1	1.093878
age	1.345121	1	1.159793

SPARTANS WILL.

Model Validation

K-fold Cross-Validation



- Used 5 random and unique data sets from original data set – Testing data sets.
- Evaluated prediction for each testing data set (k), using a linear model (with original model factors) that was generated by fitting the remaining data points of the dataset (Training data - 4/5 of the whole data set).
- Calculated R-squared (Pred) and MSE (Pred) for each iteration.

5-fold Cross-Validation

- Model Comparison

	Model 1 (Heuristic)		Model 2 (Stepwise)		Model 3 (Variable Addition)	
K	R-Squared (Pred)	MSE (Pred)	R-Squared (Pred)	MSE (Pred)	R-Squared (Pred)	MSE (Pred)
1	0.69	2.73E+10	0.759	1.98E+10	0.757	2.01E+10
2	0.699	2.55E+10	0.751	2.12E+10	0.755	2.02E+10
3	0.716	2.35E+10	0.761	2.07E+10	0.758	1.99E+10
4	0.701	2.22E+10	0.745	2.00E+10	0.756	2.10E+10
5	0.647	2.83E+10	0.752	1.93E+10	0.751	1.90E+10
Average	0.6906	2.536E+10	0.7536	2.020E+10	0.7554	2.004E+10

- Variation in R^2 (Pred) is slightly higher for Model 1, showing that the model is more sensitive to dataset variability – Lower prediction power.
- Model 3 (with less predictors) shows the least variability between R^2_k (Pred) as well as MSE_k (Pred) values. This means that this model is less susceptible to data set differences and contains higher predictive power among the three models.

SPARTANS WILL.

Final Model

Final Model

```
Call:
lm(formula = price ~ sqft_living + log(sqft_lot) + as.factor(waterfront) +
    as.factor(view) + grade + avg_total_income + age, data = final)

Residuals:
    Min      1Q  Median      3Q     Max
-861248  -81737  -10400   65837 1197515

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.887e+05  1.383e+04  -35.34   <2e-16 ***
sqft_living              1.246e+02  1.811e+00   68.82   <2e-16 ***
log(sqft_lot)           -3.354e+04  1.163e+03  -28.84   <2e-16 ***
as.factor(waterfront)1   2.546e+05  1.586e+04   16.05   <2e-16 ***
as.factor(view)1         8.086e+04  8.057e+03   10.04   <2e-16 ***
as.factor(view)2         7.913e+04  4.847e+03   16.32   <2e-16 ***
as.factor(view)3         1.521e+05  6.697e+03   22.70   <2e-16 ***
as.factor(view)4         2.519e+05  1.082e+04   23.29   <2e-16 ***
grade                    9.127e+04  1.426e+03   63.99   <2e-16 ***
avg_total_income         2.343e+03  2.311e+01  101.38   <2e-16 ***
age                      2.515e+03  3.839e+01   65.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 141900 on 21393 degrees of freedom
Multiple R-squared:  0.7544,    Adjusted R-squared:  0.7543
F-statistic:  6572 on 10 and 21393 DF,  p-value: < 2.2e-16
```

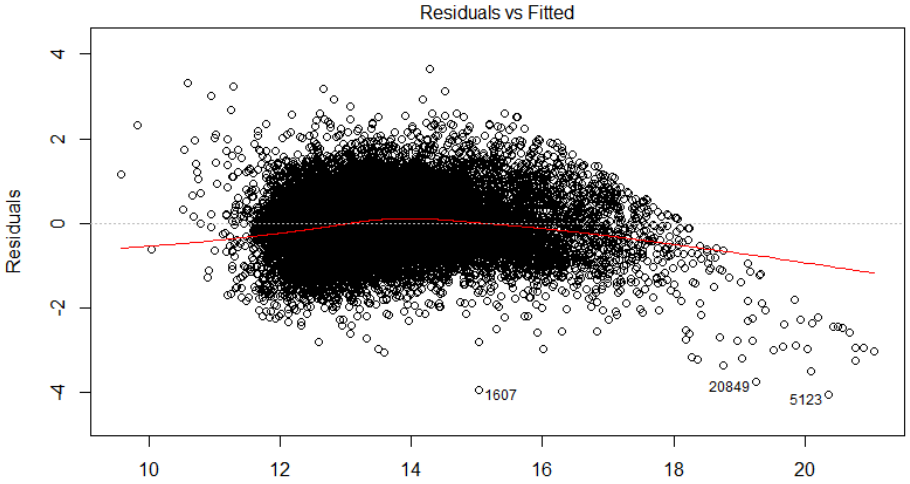
Confidence Intervals

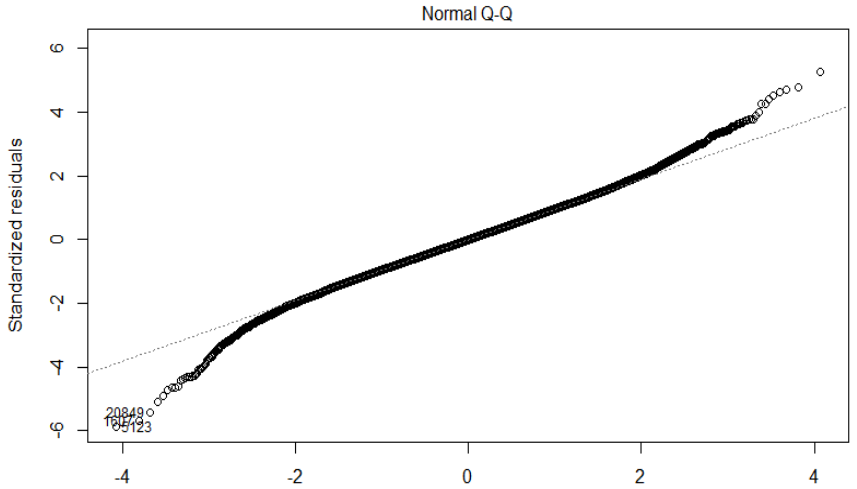
	2.5 %	97.5 %
(Intercept)	-515776.6504	-461572.3943
sqft_living	121.0792	128.1782
log(sqft_lot)	-35821.6170	-31262.0466
as.factor(waterfront)1	223502.1703	285678.7316
as.factor(view)1	65066.5892	96651.2358
as.factor(view)2	69624.2642	88627.0094
as.factor(view)3	138929.4594	165183.8260
as.factor(view)4	230694.6879	273104.1645
grade	88477.9763	94069.5823
avg_total_income	2297.3379	2387.9220
age	2439.8413	2590.3361

No multicollinearity

	GVIF	Df	GVIF^(1/(2*Df))
sqft_living	2.592269	1	1.610052
log(sqft_lot)	1.167033	1	1.080293
as.factor(waterfront)	1.415895	1	1.189914
as.factor(view)	1.560521	4	1.057204
grade	2.792675	1	1.671130
avg_total_income	1.196570	1	1.093878
age	1.345121	1	1.159793

Final Model - Diagnostics





SPARTANS WILL.

Takeaways

Takeaways

- The most important variable were grade and income: which is an indication of the quality of construction and affordability
- Variable addition was the most optimal model and produced the best diagnostics
- Linear regression was determined to be an average technique to explain the house prices
- It was interesting to see the minimal impact of bedroom and sqft_lot on the price
- Slight non-constant variance could be observed; this was mostly due to the houses priced over $1 million

SPARTANS WILL.

Appendix

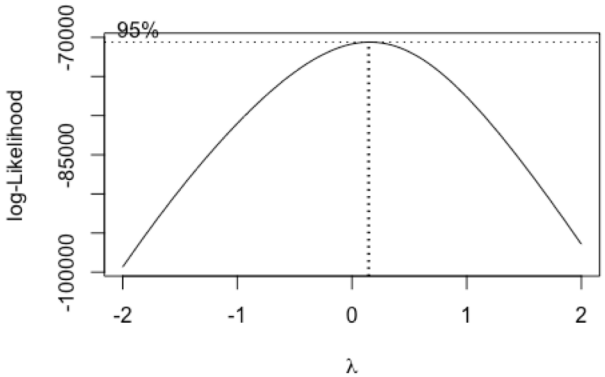
Location! Location! Location!

- Create a model with zipcode as a categorical variable since house prices are all about location

- Disregard model due to concerns with overfitting and complication (too many variables)

Model Statistics

```
model<-
lm(price~bathrooms+sqft_living+sqft_lot+waterfront+view+condition+
        grade+age+zipcode+bedrooms, data=house)
```



Transform price to power of 0.15

Model Diagnostics

```
Residual standard error: 0.1964 on 21323 degrees of freedom
Multiple R-squared:  0.8658,    Adjusted R-squared:  0.8653
F-statistic:  1698 on 81 and 21323 DF,  p-value: < 2.2e-16
```