Project 2
House Sales in King County, USA

Spencer Stucky, Cheng Zeng, Wenyu Zeng, Chao Zhou
12/1/2019

Agenda

Introduction of Dataset

Project Objective

ANOVA

Linear Regression

PCA/PCR

K-fold Cross Validation

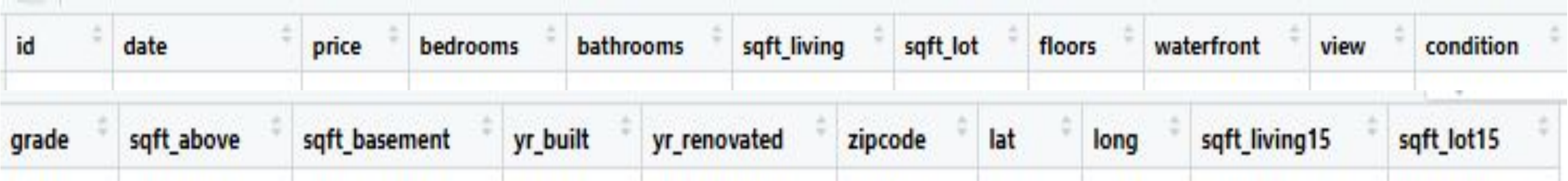
Next Step…

Introduction of Dataset

- House Sales of King County, Washington state

- Download from Kaggle, provided by King County

- Includes homes sold between 2014 to 2015, total 21613 observations and 21 features



id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition

grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15

View - An index from 0 to 4 of how good the view of the property was

Grade - An index from 1 to 13, where grading from short of building construction and design to high quality level of construction and design

Project Objective

What features relate to the price of house

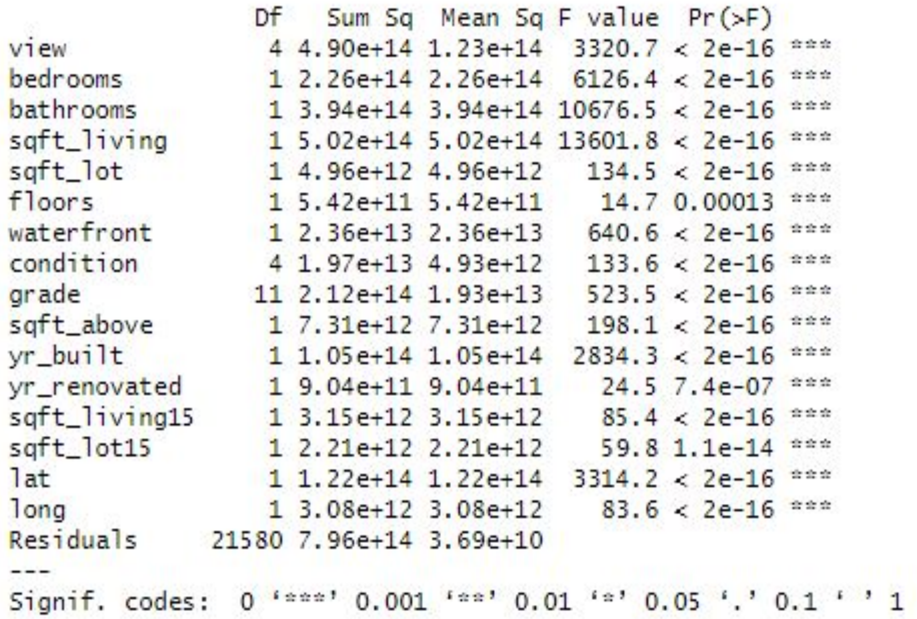
Predict the price of housing of King County, USA

ANOVA

ANOVA Test

- **Null Hypothesis**:
  - Prices of houses with different features are equal.

- **Alternative hypothesis**:
  - Prices of houses with different features are not equal.

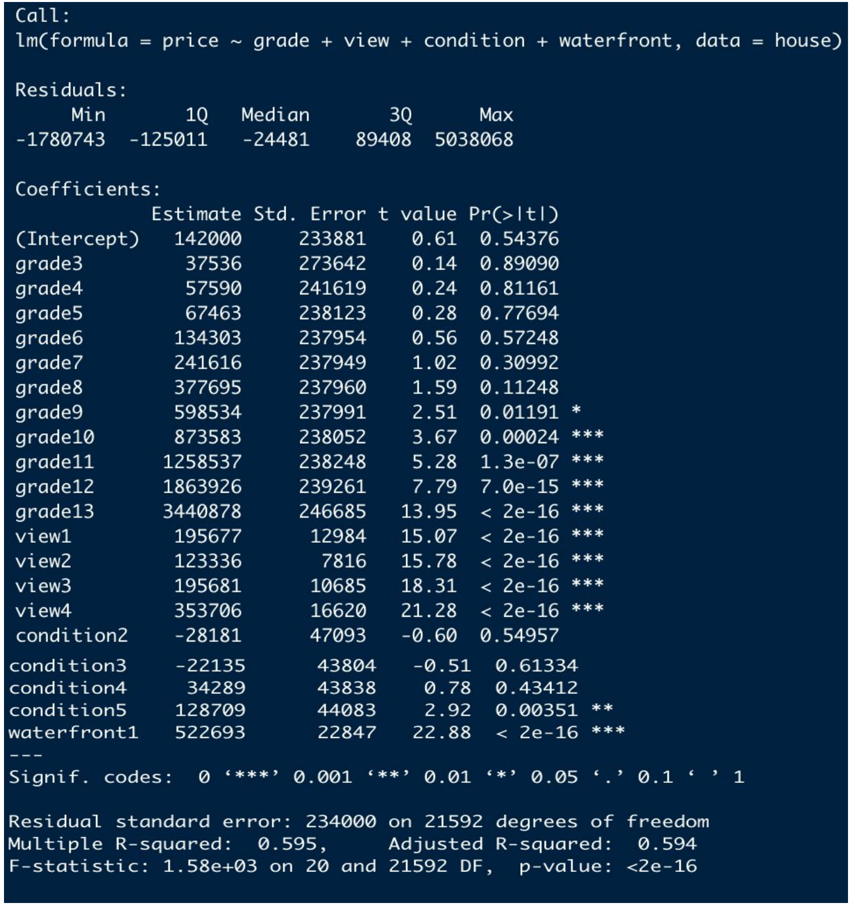


```
                 Df   Sum Sq  Mean Sq F value   Pr(>F)    
view              4 4.90e+14 1.23e+14  3320.7  < 2e-16 ***
bedrooms          1 2.26e+14 2.26e+14  6126.4  < 2e-16 ***
bathrooms         1 3.94e+14 3.94e+14 10676.5  < 2e-16 ***
sqft_living       1 5.02e+14 5.02e+14 13601.8  < 2e-16 ***
sqft_lot          1 4.96e+12 4.96e+12   134.5  < 2e-16 ***
floors            1 5.42e+11 5.42e+11    14.7  0.00013 ***
waterfront        1 2.36e+13 2.36e+13   640.6  < 2e-16 ***
condition         4 1.97e+13 4.93e+12   133.6  < 2e-16 ***
grade            11 2.12e+14 1.93e+13   523.5  < 2e-16 ***
sqft_above        1 7.31e+12 7.31e+12   198.1  < 2e-16 ***
yr_built          1 1.05e+14 1.05e+14  2834.3  < 2e-16 ***
yr_renovated      1 9.04e+11 9.04e+11    24.5  7.4e-07 ***
sqft_living15     1 3.15e+12 3.15e+12    85.4  < 2e-16 ***
sqft_lot15        1 2.21e+12 2.21e+12    59.8  1.1e-14 ***
lat               1 1.22e+14 1.22e+14  3314.2  < 2e-16 ***
long              1 3.08e+12 3.08e+12    83.6  < 2e-16 ***
Residuals     21580 7.96e+14 3.69e+10                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear Regression

Linear Regression M1

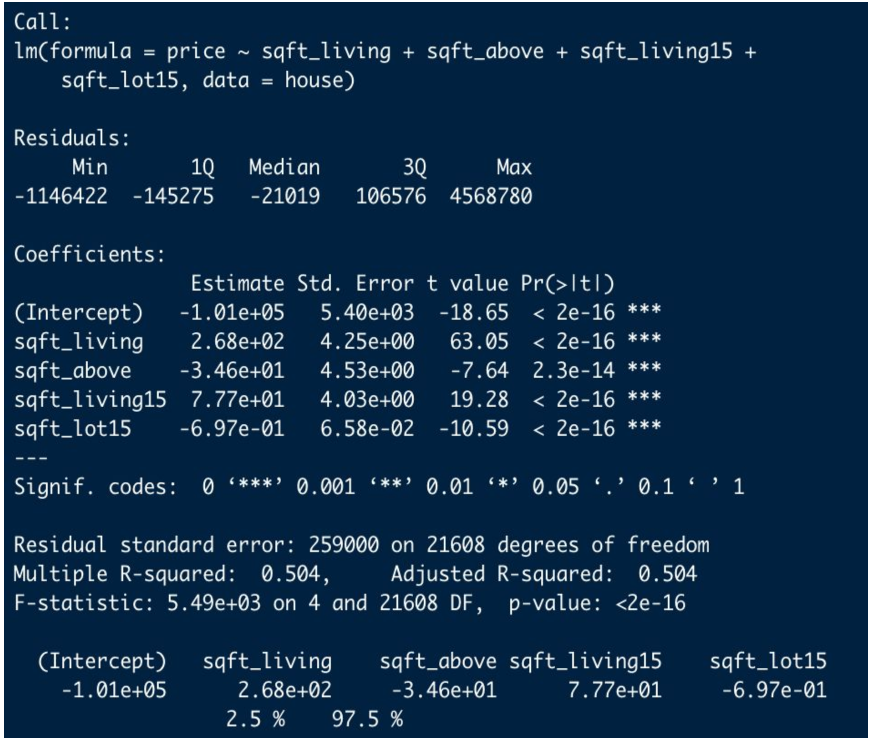
- In this **model**, we first wanted to look at the categorical variables for house attributes and ranking.
- R2 is **59**% thus model explains 59% of price data.
- Grade ratings 9-13 are sig as well as higher-end of condition of house. View and waterfront are sig
- All categorical variables are positively **correlated** with price and thus drive up price as they increase.



```
Call:
lm(formula = price ~ grade + view + condition + waterfront, data = house)

Residuals:
     Min       1Q   Median       3Q      Max 
-1780743  -125011   -24481    89408  5038068 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   142000     233881    0.61  0.54376    
grade3         37536     273642    0.14  0.89090    
grade4         57590     241619    0.24  0.81161    
grade5         67463     238123    0.28  0.77694    
grade6        134303     237954    0.56  0.57248    
grade7        241616     237949    1.02  0.30992    
grade8        377695     237960    1.59  0.11248    
grade9        598534     237991    2.51  0.01191 *  
grade10       873583     238052    3.67  0.00024 ***
grade11      1258537     238248    5.28  1.3e-07 ***
grade12      1863926     239261    7.79  7.0e-15 ***
grade13      3440878     246685   13.95  < 2e-16 ***
view1         195677      12984   15.07  < 2e-16 ***
view2         123336       7816   15.78  < 2e-16 ***
view3         195681      10685   18.31  < 2e-16 ***
view4         353706      16620   21.28  < 2e-16 ***
condition2    -28181      47093   -0.60  0.54957    
condition3    -22135      43804   -0.51  0.61334    
condition4     34289      43838    0.78  0.43412    
condition5    128709      44083    2.92  0.00351 ** 
waterfront1   522693      22847   22.88  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 234000 on 21592 degrees of freedom
Multiple R-squared:  0.595,	Adjusted R-squared:  0.594 
F-statistic: 1.58e+03 on 20 and 21592 DF,  p-value: <2e-16
```

Linear Regression M2

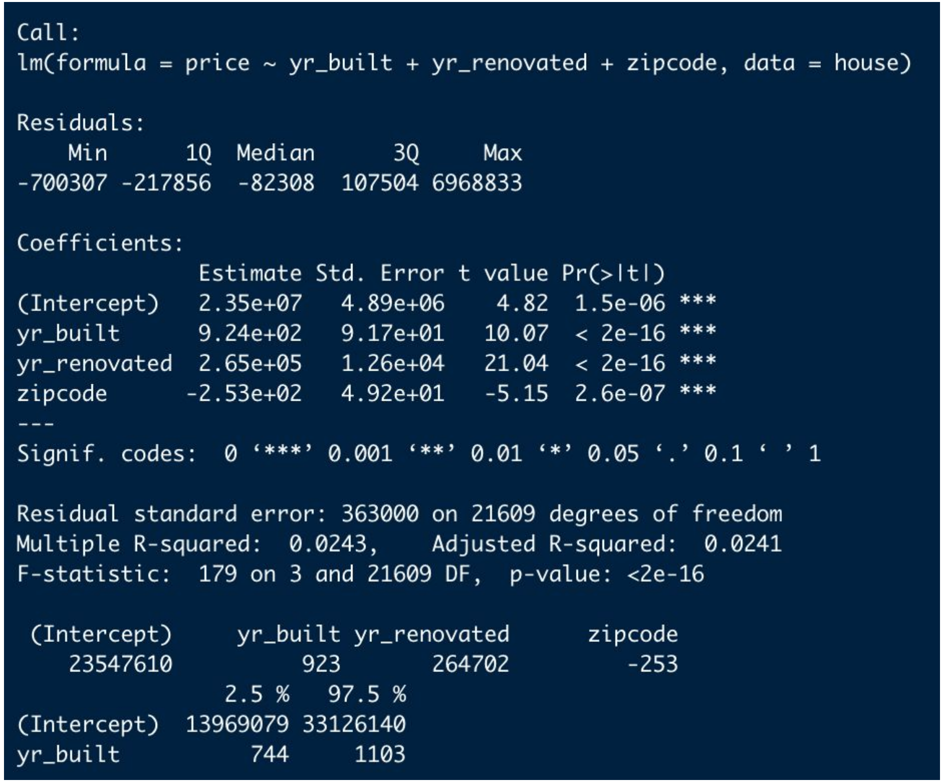
- Idea behind this **model** was to look at sq footage, size of house and surrounding neighborhood to determine if they effect price in some way. Also to look at how the sq ft variables may be collinear
- R2 is **50%**
- All p values are significant.
- Sqft living and Sq ft living 15 are positively **correlated** with price while sqft above and sqft lot 15 are negatively correlated with price.
- Some moderately high VIFs for sqft living and sqft above - moderate collinearity there.



```
Call:
lm(formula = price ~ sqft_living + sqft_above + sqft_living15 + 
    sqft_lot15, data = house)

Residuals:
     Min       1Q   Median       3Q      Max 
-1146422  -145275   -21019   106576  4568780 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.01e+05   5.40e+03  -18.65  < 2e-16 ***
sqft_living    2.68e+02   4.25e+00   63.05  < 2e-16 ***
sqft_above    -3.46e+01   4.53e+00   -7.64  2.3e-14 ***
sqft_living15  7.77e+01   4.03e+00   19.28  < 2e-16 ***
sqft_lot15    -6.97e-01   6.58e-02  -10.59  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 259000 on 21608 degrees of freedom
Multiple R-squared:  0.504,	Adjusted R-squared:  0.504 
F-statistic: 5.49e+03 on 4 and 21608 DF,  p-value: <2e-16

  (Intercept)   sqft_living    sqft_above sqft_living15    sqft_lot15 
    -1.01e+05      2.68e+02     -3.46e+01      7.77e+01     -6.97e-01 
               2.5 %    97.5 %
```

Linear Regression M3

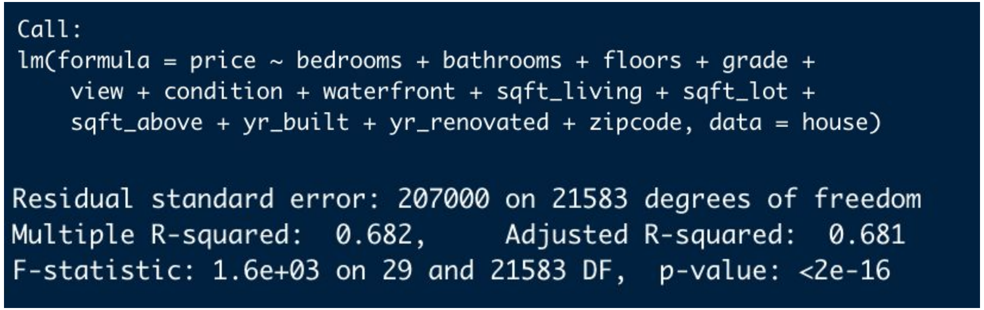
- Idea behind this **model** was to look at yr built, yr renovated, and location of housing to see if they determined something about price.
- All are highly significant but R2 is very low at **2.5%** thus model does not do a good job of explaining results in data
- Year built and Yr renovated were positively correlated with price, with yr built increasing price by 924.



```
Call:
lm(formula = price ~ yr_built + yr_renovated + zipcode, data = house)

Residuals:
    Min      1Q  Median      3Q     Max 
-700307 -217856  -82308  107504 6968833 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.35e+07   4.89e+06    4.82  1.5e-06 ***
yr_built      9.24e+02   9.17e+01   10.07  < 2e-16 ***
yr_renovated  2.65e+05   1.26e+04   21.04  < 2e-16 ***
zipcode      -2.53e+02   4.92e+01   -5.15  2.6e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 363000 on 21609 degrees of freedom
Multiple R-squared:  0.0243,	Adjusted R-squared:  0.0241 
F-statistic:  179 on 3 and 21609 DF,  p-value: <2e-16

 (Intercept)     yr_built yr_renovated      zipcode 
    23547610          923       264702         -253 
                2.5 %   97.5 %
(Intercept)  13969079 33126140
yr_built          744     1103
```

Linear Regression M4

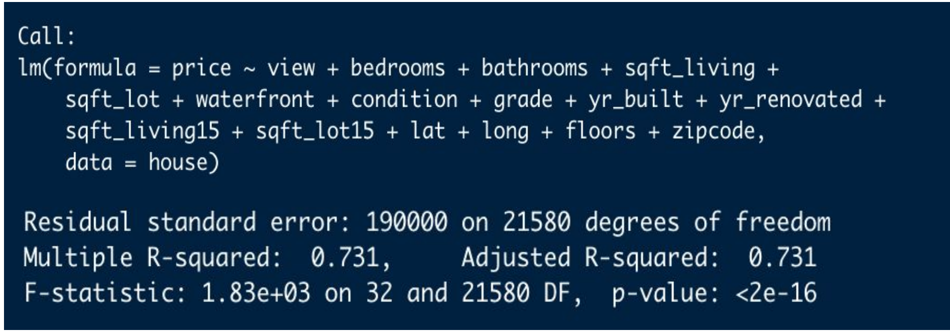
- This **model** used all variables except the variables on surrounding houses and sq basement
- R2 is **68%**.
- Bedrooms, sqft lot, sqft above, yr built are negatively **correlated** with price.
- High VIFs for sqft living and sqft above indicate multicollinearity but nothing too extreme
- All p values are highly significant at the .001 level



```
Call:
lm(formula = price ~ bedrooms + bathrooms + floors + grade +
    view + condition + waterfront + sqft_living + sqft_lot +
    sqft_above + yr_built + yr_renovated + zipcode, data = house)

Residual standard error: 207000 on 21583 degrees of freedom
Multiple R-squared:  0.682,     Adjusted R-squared:  0.681
F-statistic: 1.6e+03 on 29 and 21583 DF,  p-value: <2e-16
```

Linear Regression M5

- Started with **model** of all variables, then removed sqft above, sqft basement (not sig)
- R2 is about **73%**
- As bedrooms increase price decreases, yr built, sq ft of surrounding lots, and zipcode are also negatively correlated with price
- A moderately high VIF of 5 for sqft living - not high enough for concern
- All p values are highly significant at the .001 level



```
Call:
lm(formula = price ~ view + bedrooms + bathrooms + sqft_living +
    sqft_lot + waterfront + condition + grade + yr_built + yr_renovated +
    sqft_living15 + sqft_lot15 + lat + long + floors + zipcode,
    data = house)

Residual standard error: 190000 on 21580 degrees of freedom
Multiple R-squared:  0.731,     Adjusted R-squared:  0.731
F-statistic: 1.83e+03 on 32 and 21580 DF,  p-value: <2e-16
```

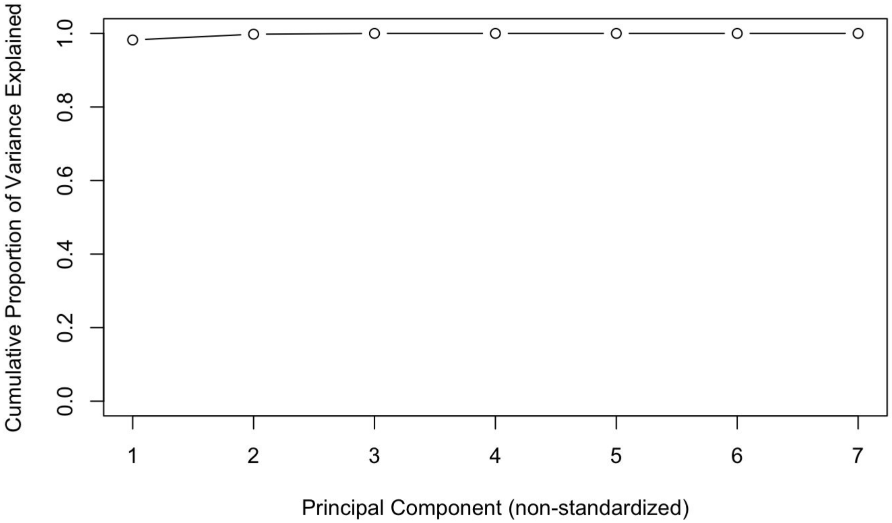
Linear Regression Findings on Price

- # of bedrooms were consistently negatively correlated with housing price
- Bathrooms were consistently positively correlated with housing price
- Waterfront, year renovated, and higher-end ratings of condition had effects on price upwards of $500+
- Grade ratings of 11-13 gave significant boosts to housing prices
- M5 returned highest R2 w/ significance

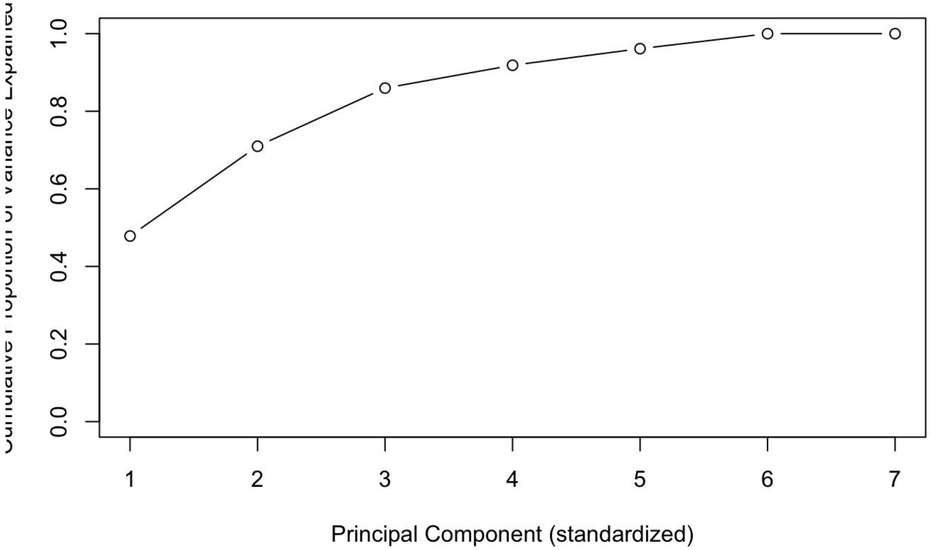
PCA/PCR

PCA without Factor

❖ Unscaled

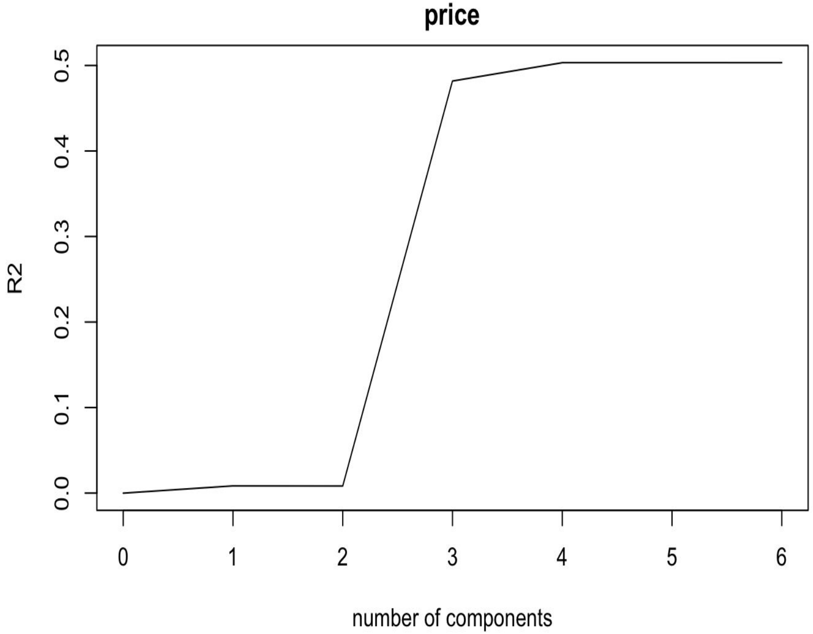


❖ Scaled

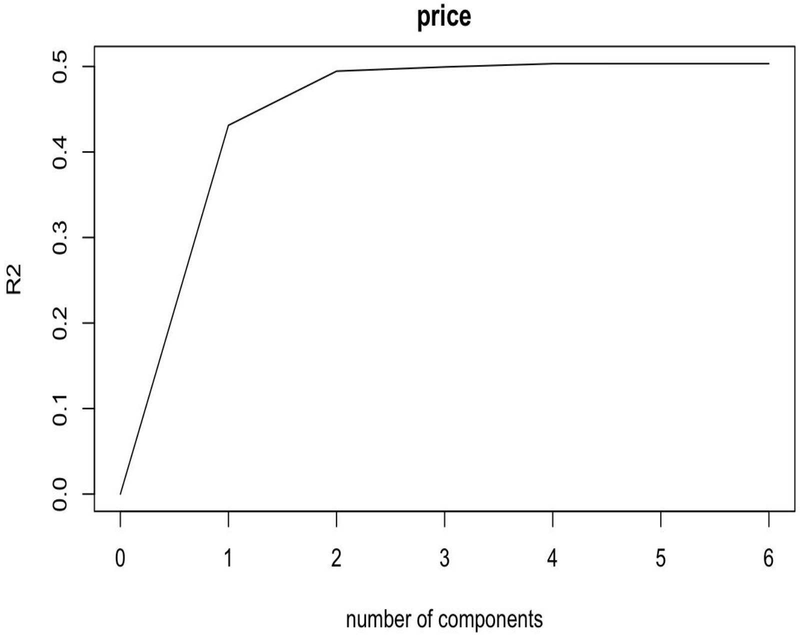


PCR without Factor

- Unscaled



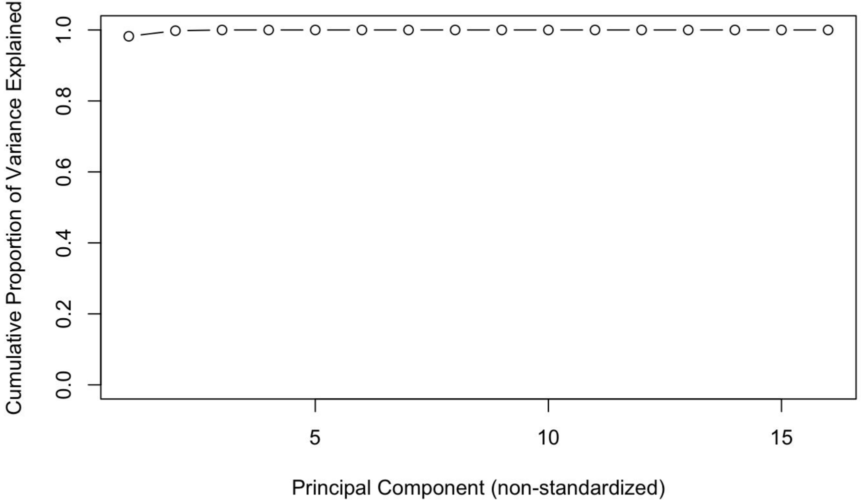
- Scaled

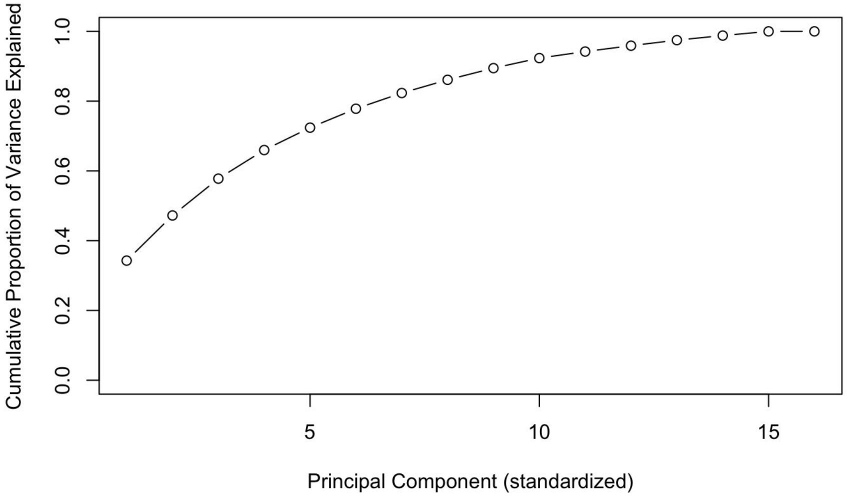


PCA with Factor

- Unscaled

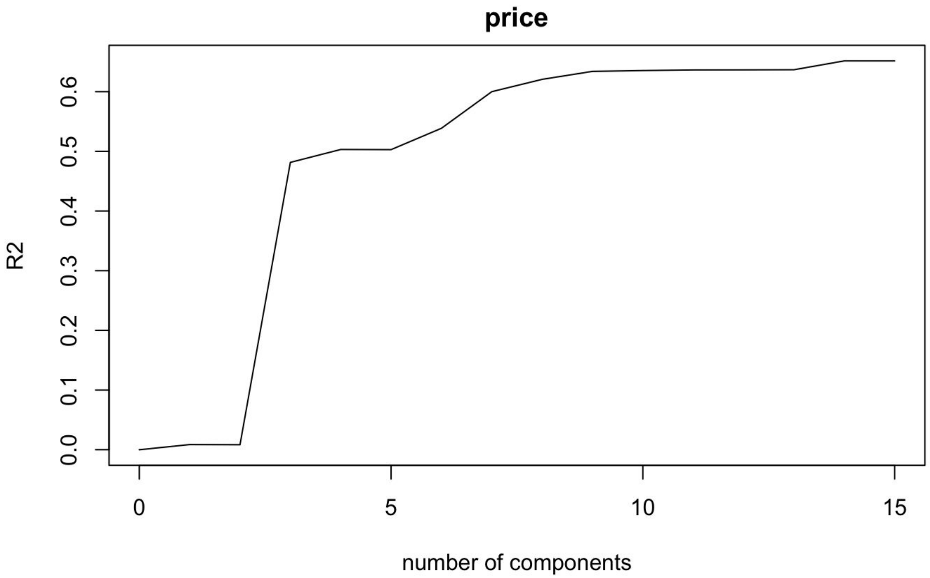
- Scaled



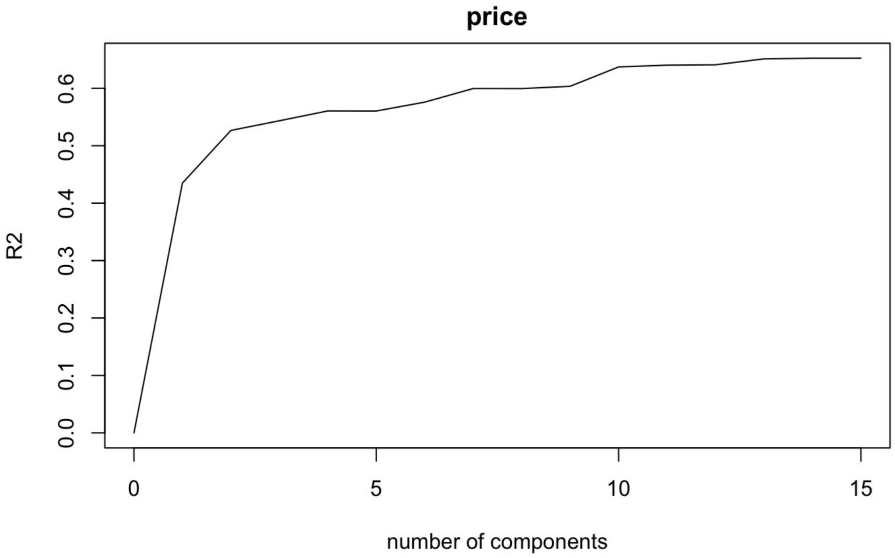


PCR with Factor

- Unscaled

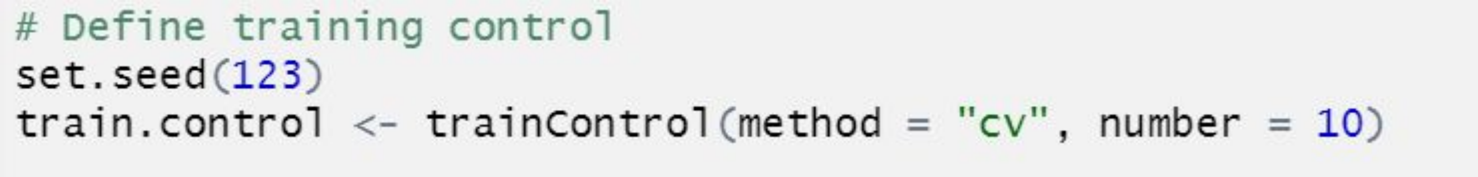


- Scaled



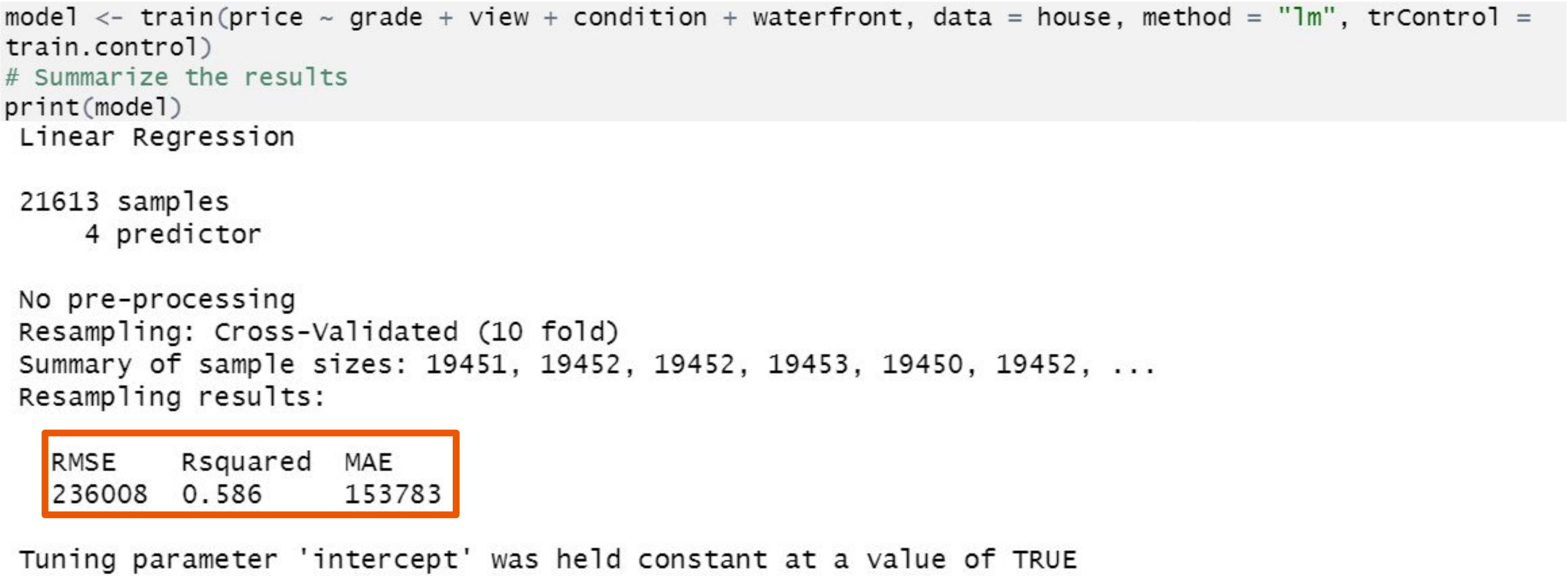
K-fold Cross Validation

Setting

- Used Package “caret”
- 10 fold cross validation (K=10)
- 

```
# Define training control
set.seed(123)
train.control <- trainControl(method = "cv", number = 10)
```

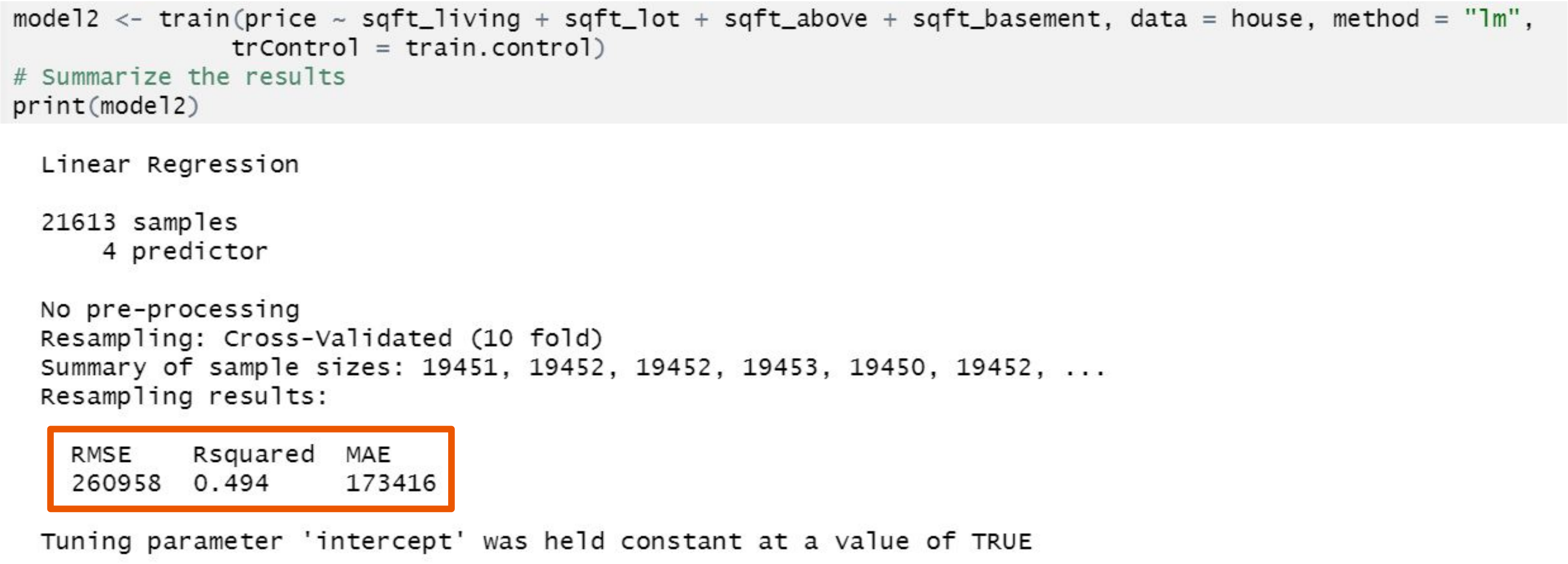
Cross Validation on Model 1



```r
model <- train(price ~ grade + view + condition + waterfront, data = house, method = "lm", trControl =
train.control)
# Summarize the results
print(model)
```

```
Linear Regression

21613 samples
    4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results:

  RMSE     Rsquared  MAE
  236008   0.586     153783

Tuning parameter 'intercept' was held constant at a value of TRUE
```

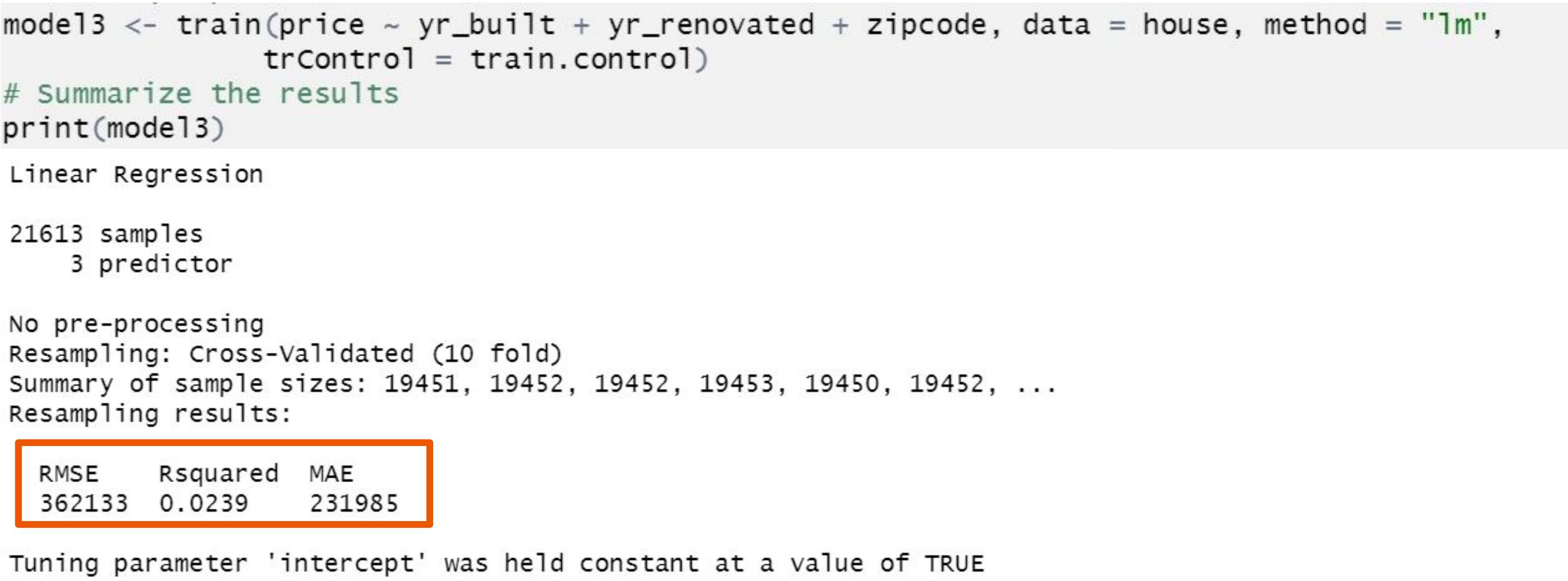
Cross Validation on Model 2



```
model2 <- train(price ~ sqft_living + sqft_lot + sqft_above + sqft_basement, data = house, method = "lm",
               trControl = train.control)
# Summarize the results
print(model2)
```

```
  Linear Regression 

  21613 samples
      4 predictor

  No pre-processing
  Resampling: Cross-Validated (10 fold) 
  Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ... 
  Resampling results:

    RMSE    Rsquared   MAE   
    260958  0.494      173416

  Tuning parameter 'intercept' was held constant at a value of TRUE
```

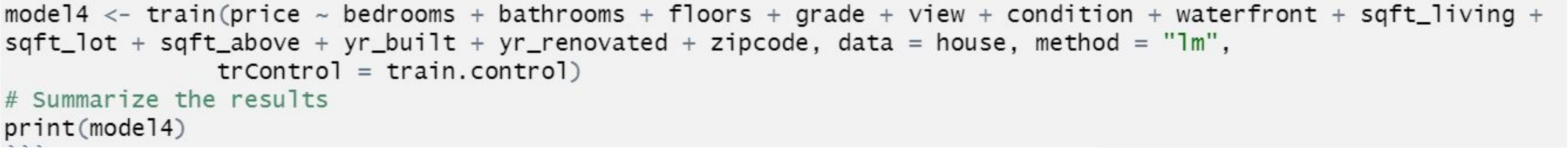
Cross Validation on Model 3



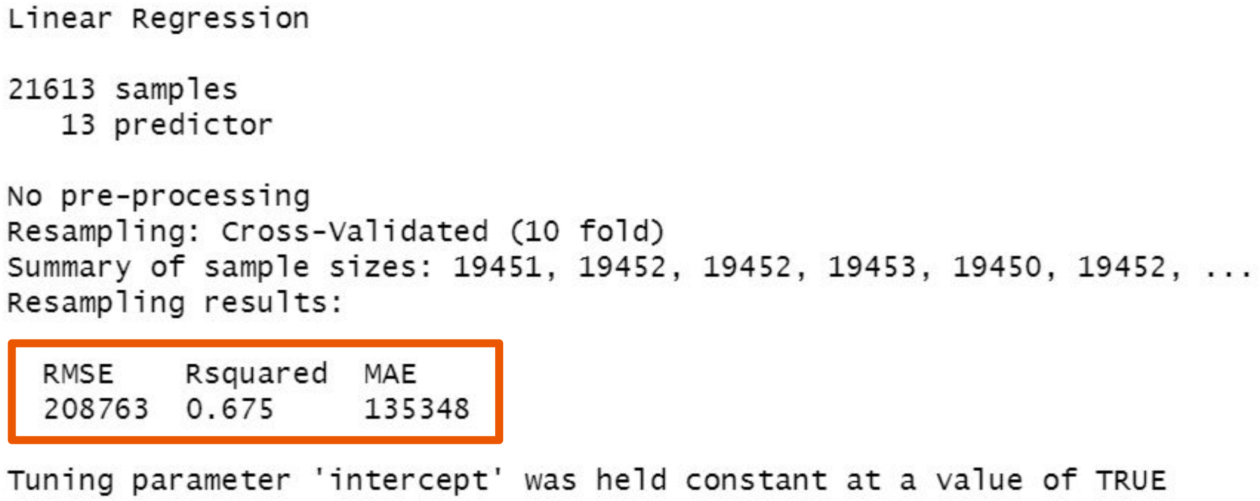
```
model3 <- train(price ~ yr_built + yr_renovated + zipcode, data = house, method = "lm",
               trControl = train.control)
# Summarize the results
print(model3)
```

```
Linear Regression

21613 samples
    3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results:

  RMSE     Rsquared   MAE
  362133   0.0239     231985

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Cross Validation on Model 4

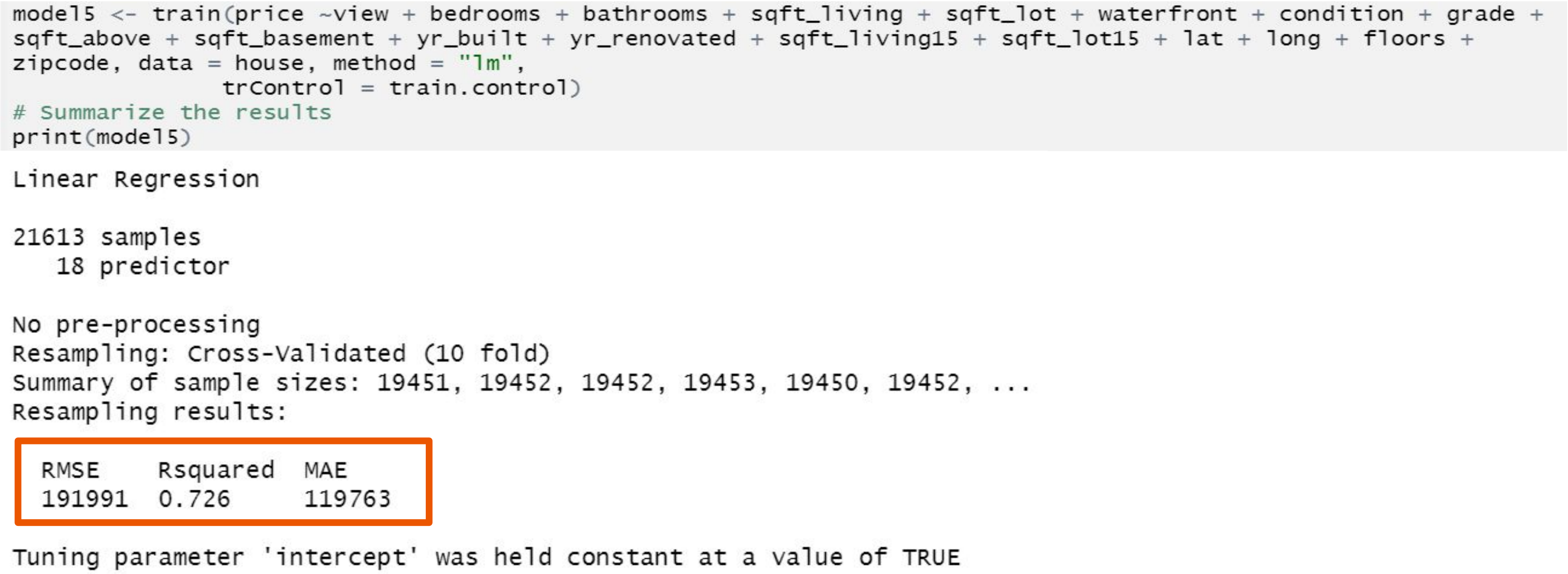


```
model4 <- train(price ~ bedrooms + bathrooms + floors + grade + view + condition + waterfront + sqft_living +
sqft_lot + sqft_above + yr_built + yr_renovated + zipcode, data = house, method = "lm",
               trControl = train.control)
# Summarize the results
print(model4)
```



```
Linear Regression

21613 samples
   13 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results:

  RMSE      Rsquared   MAE
  208763    0.675      135348

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Cross-Validation on Full Model

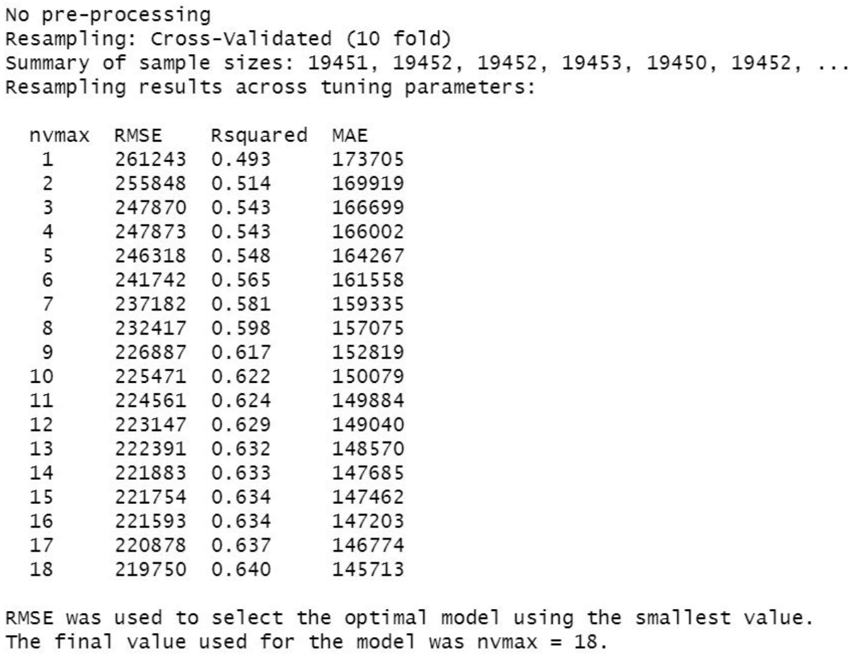


```
model5 <- train(price ~view + bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + condition + grade +
sqft_above + sqft_basement + yr_built + yr_renovated + sqft_living15 + sqft_lot15 + lat + long + floors +
zipcode, data = house, method = "lm",
               trControl = train.control)
# Summarize the results
print(model5)
```

```
Linear Regression

21613 samples
   18 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results:

  RMSE    Rsquared  MAE
  191991  0.726     119763

Tuning parameter 'intercept' was held constant at a value of TRUE
```

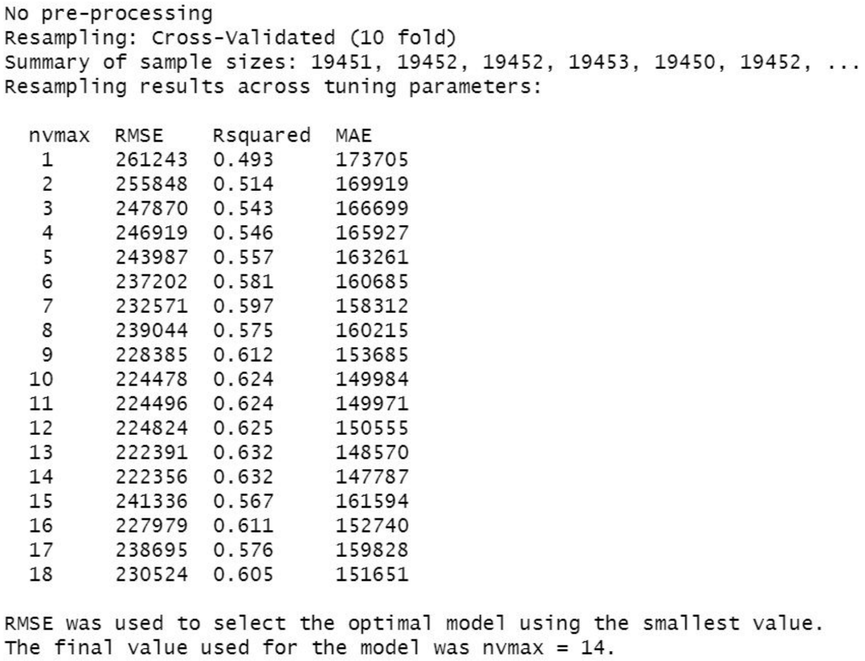
Forward/Stepwise Selection with Full Model

- Forward Selection



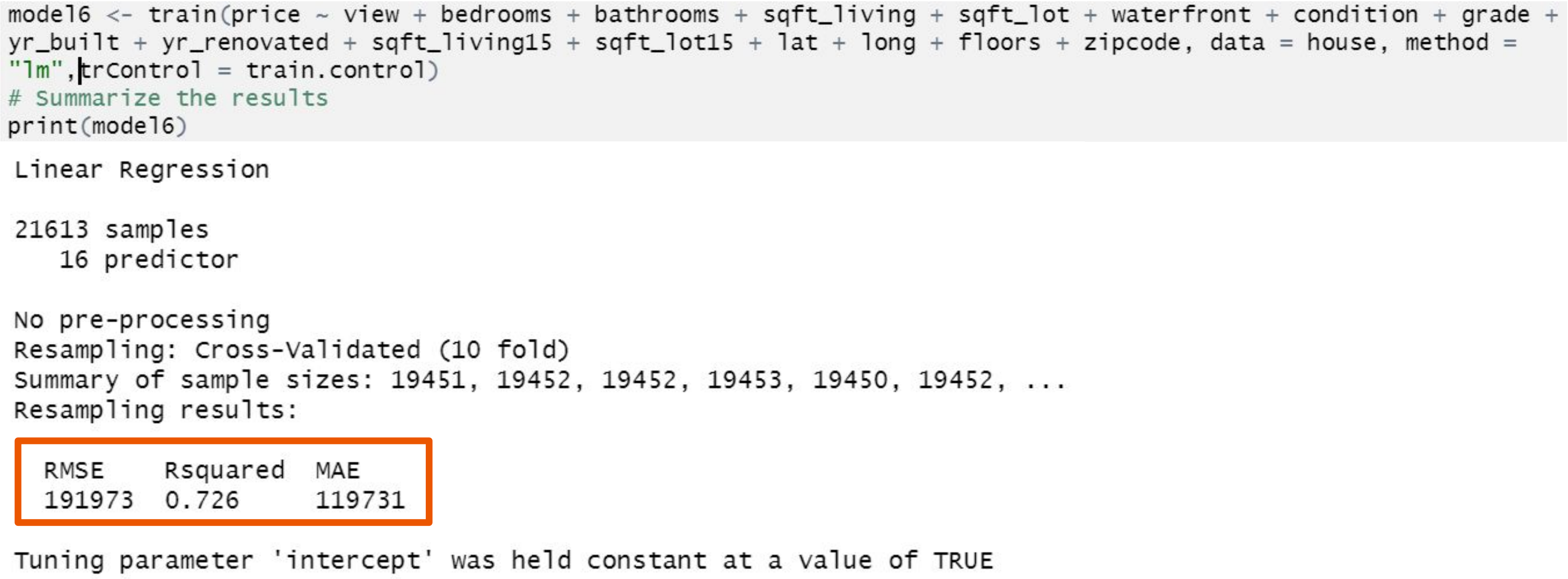
```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results across tuning parameters:

  nvmax  RMSE    Rsquared  MAE
   1     261243  0.493     173705
   2     255848  0.514     169919
   3     247870  0.543     166699
   4     247873  0.543     166002
   5     246318  0.548     164267
   6     241742  0.565     161558
   7     237182  0.581     159335
   8     232417  0.598     157075
   9     226887  0.617     152819
  10     225471  0.622     150079
  11     224561  0.624     149884
  12     223147  0.629     149040
  13     222391  0.632     148570
  14     221883  0.633     147685
  15     221754  0.634     147462
  16     221593  0.634     147203
  17     220878  0.637     146774
  18     219750  0.640     145713

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was nvmax = 18.
```

- Stepwise Selection



```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results across tuning parameters:

  nvmax  RMSE    Rsquared  MAE
   1     261243  0.493     173705
   2     255848  0.514     169919
   3     247870  0.543     166699
   4     246919  0.546     165927
   5     243987  0.557     163261
   6     237202  0.581     160685
   7     232571  0.597     158312
   8     239044  0.575     160215
   9     228385  0.612     153685
  10     224478  0.624     149984
  11     224496  0.624     149971
  12     224824  0.625     150555
  13     222391  0.632     148570
  14     222356  0.632     147787
  15     241336  0.567     161594
  16     227979  0.611     152740
  17     238695  0.576     159828
  18     230524  0.605     151651

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was nvmax = 14.
```

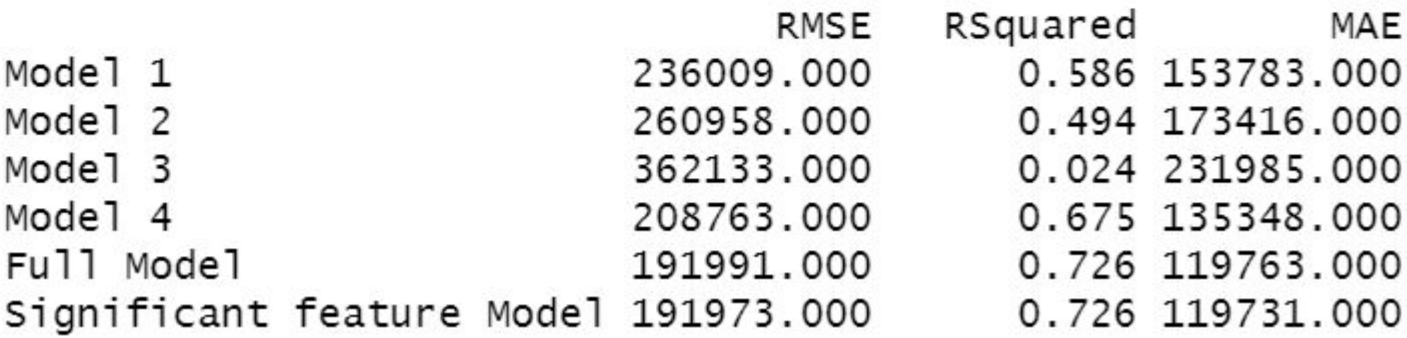
Cross-Validation on all significant factors



```
model6 <- train(price ~ view + bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + condition + grade +
yr_built + yr_renovated + sqft_living15 + sqft_lot15 + lat + long + floors + zipcode, data = house, method =
"lm",trControl = train.control)
# Summarize the results
print(model6)
```

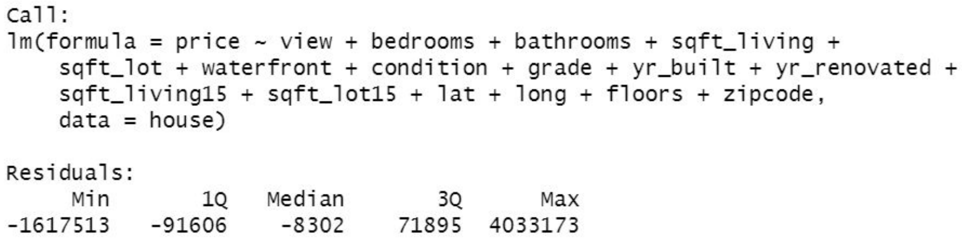
```
Linear Regression

21613 samples
   16 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 19451, 19452, 19452, 19453, 19450, 19452, ...
Resampling results:

  RMSE    Rsquared  MAE
  191973  0.726     119731

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Model Comparison

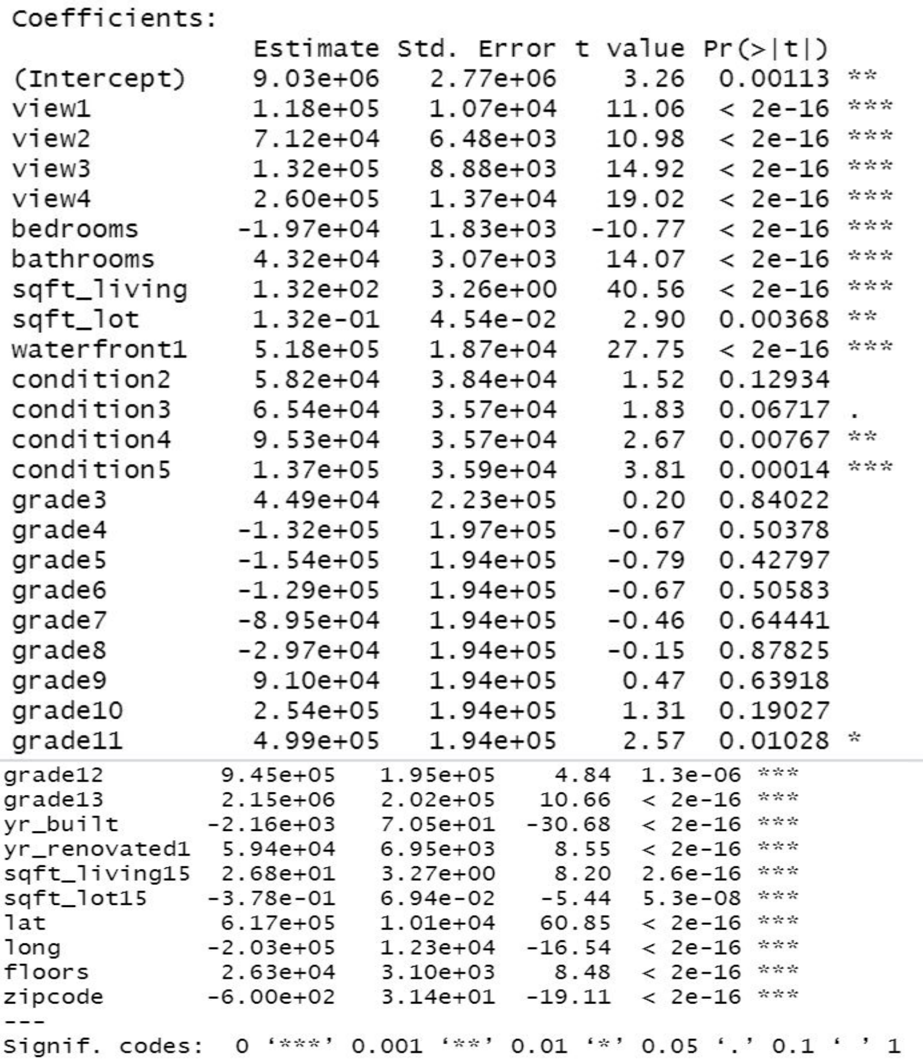


```
                              RMSE    RSquared        MAE
Model 1                     236009.000     0.586 153783.000
Model 2                     260958.000     0.494 173416.000
Model 3                     362133.000     0.024 231985.000
Model 4                     208763.000     0.675 135348.000
Full Model                  191991.000     0.726 119763.000
Significant feature Model 191973.000     0.726 119731.000
```

Best Model so far



```
Call:
lm(formula = price ~ view + bedrooms + bathrooms + sqft_living +
    sqft_lot + waterfront + condition + grade + yr_built + yr_renovated +
    sqft_living15 + sqft_lot15 + lat + long + floors + zipcode,
    data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-1617513   -91606    -8302    71895  4033173
```



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.03e+06   2.77e+06    3.26  0.00113 **
view1          1.18e+05   1.07e+04   11.06  < 2e-16 ***
view2          7.12e+04   6.48e+03   10.98  < 2e-16 ***
view3          1.32e+05   8.88e+03   14.92  < 2e-16 ***
view4          2.60e+05   1.37e+04   19.02  < 2e-16 ***
bedrooms      -1.97e+04   1.83e+03  -10.77  < 2e-16 ***
bathrooms      4.32e+04   3.07e+03   14.07  < 2e-16 ***
sqft_living    1.32e+02   3.26e+00   40.56  < 2e-16 ***
sqft_lot       1.32e-01   4.54e-02    2.90  0.00368 **
waterfront1    5.18e+05   1.87e+04   27.75  < 2e-16 ***
condition2     5.82e+04   3.84e+04    1.52  0.12934
condition3     6.54e+04   3.57e+04    1.83  0.06717 .
condition4     9.53e+04   3.57e+04    2.67  0.00767 **
condition5     1.37e+05   3.59e+04    3.81  0.00014 ***
grade3         4.49e+04   2.23e+05    0.20  0.84022
grade4        -1.32e+05   1.97e+05   -0.67  0.50378
grade5        -1.54e+05   1.94e+05   -0.79  0.42797
grade6        -1.29e+05   1.94e+05   -0.67  0.50583
grade7        -8.95e+04   1.94e+05   -0.46  0.64441
grade8        -2.97e+04   1.94e+05   -0.15  0.87825
grade9         9.10e+04   1.94e+05    0.47  0.63918
grade10        2.54e+05   1.94e+05    1.31  0.19027
grade11        4.99e+05   1.94e+05    2.57  0.01028 *
grade12        9.45e+05   1.95e+05    4.84  1.3e-06 ***
grade13        2.15e+06   2.02e+05   10.66  < 2e-16 ***
yr_built      -2.16e+03   7.05e+01  -30.68  < 2e-16 ***
yr_renovated1  5.94e+04   6.95e+03    8.55  < 2e-16 ***
sqft_living15  2.68e+01   3.27e+00    8.20  2.6e-16 ***
sqft_lot15    -3.78e-01   6.94e-02   -5.44  5.3e-08 ***
lat            6.17e+05   1.01e+04   60.85  < 2e-16 ***
long          -2.03e+05   1.23e+04  -16.54  < 2e-16 ***
floors         2.63e+04   3.10e+03    8.48  < 2e-16 ***
zipcode       -6.00e+02   3.14e+01  -19.11  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Next Step...

- Try other Regression Models
- ...

Thank you!