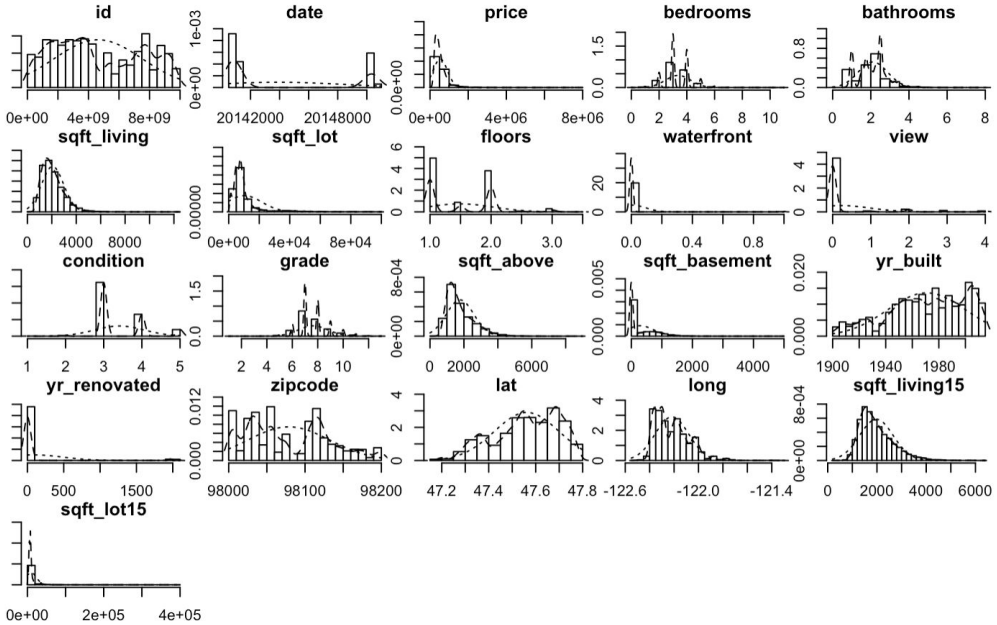
HOUSING PRICE PREDICTION

Sirui Li
May 7, 2020

DATASET

Dataset

- House sale prices for King County, WA from May 2014 to May 2015
- Scraped from KingCounty.gov
- Dimension: 21 columns and 21,613 rows
- Data cleaning & transforming: remove outliers and log transformation



MULTIPLE REGRESSION: LOCATION LOCATION LOCATION

Location Matters

Compare models 1 and reduced model without location related features.

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  21602 1948.3
2  21605 2952.2 -3   -1003.9 3710.4 < 2.2e-16 ***
```

Why?

Good school
Safe neighborhood
Facilities around
Public transportation
Views and waterfront
...

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     -8.763e+01  4.097e+00 -21.388  < 2e-16 ***
log(sqft_living)                                 1.199e-01  2.393e-02   5.011 5.47e-07 ***
log(sqft_lot)                                    2.191e-02  5.809e-03   3.773 0.000162 ***
ifelse(sqft_basement == 0, 0, log(sqft_basement))  3.093e-02  1.629e-03  18.988  < 2e-16 ***
log(sqft_living15)                               3.956e-01  9.989e-03  39.606  < 2e-16 ***
log(sqft_lot15)                                 -3.525e-02  6.449e-03  -5.467 4.64e-08 ***
log(sqft_above)                                  5.024e-01  2.292e-02  21.923  < 2e-16 ***
zipcode                                         -3.742e-04  4.818e-05  -7.766 8.44e-15 ***
long                                            -4.584e-01  1.897e-02 -24.161  < 2e-16 ***
lat                                              1.552e+00  1.563e-02  99.298  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3003 on 21602 degrees of freedom
Multiple R-squared:  0.675,	Adjusted R-squared:  0.6749
F-statistic:  4985 on 9 and 21602 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      5.805451   0.058740  98.833  < 2e-16 ***
log(sqft_living)                                 0.073756   0.029445   2.505   0.0123 *
log(sqft_lot)                                   -0.014708   0.007128  -2.063   0.0391 *
ifelse(sqft_basement == 0, 0, log(sqft_basement))  0.045074   0.001988  22.673  < 2e-16 ***
log(sqft_living15)                               0.451344   0.012098  37.308  < 2e-16 ***
log(sqft_lot15)                                 -0.061537   0.007920  -7.770  8.2e-15 ***
log(sqft_above)                                  0.520939   0.028207  18.469  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3697 on 21605 degrees of freedom
Multiple R-squared:  0.5076,	Adjusted R-squared:  0.5074
F-statistic:  3711 on 6 and 21605 DF,  p-value: < 2.2e-16
```

MORE ON MULTIPLE REGRESSION

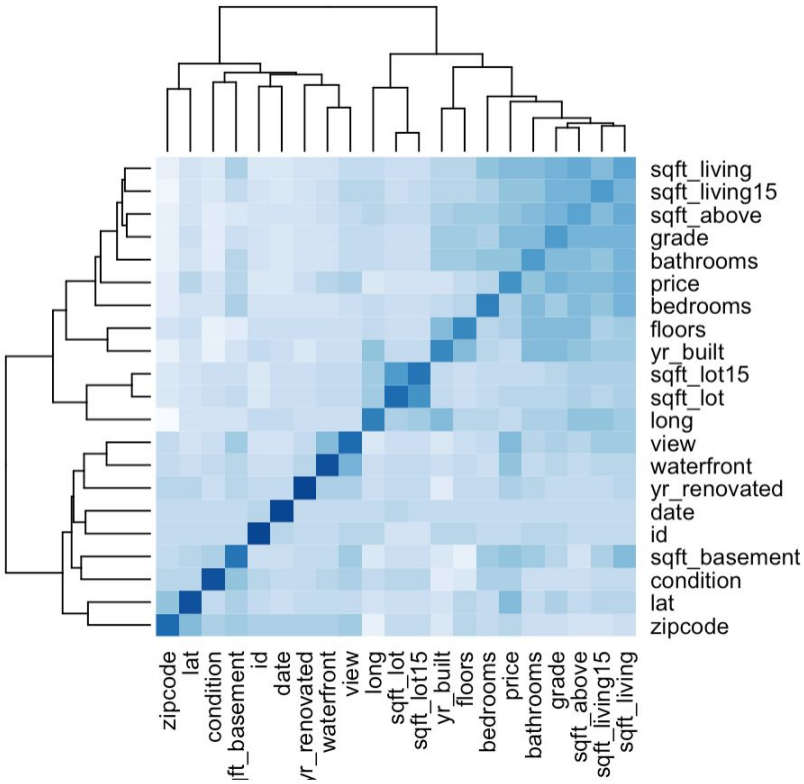
Variables Removed

Selectively remove parameters with high correlation and p-value > 0.05 (4 removed)

Full model vs. Reduced

```
Residual standard error: 0.2465 on 21126 degrees of freedom
Multiple R-squared:  0.781,  Adjusted R-squared:  0.7808
F-statistic:  3965 on 19 and 21126 DF,  p-value: < 2.2e-16

Residual standard error: 0.2849 on 21130 degrees of freedom
Multiple R-squared:  0.7076, Adjusted R-squared:  0.7074
F-statistic:  3409 on 15 and 21130 DF,  p-value: < 2.2e-16
```



POLYNOMIAL REGRESSION

Polynomial Regression Models

Full model with degree = 2 vs. Reduced Model

```
Residual standard error: 0.2306 on 21110 degrees of freedom
Multiple R-squared:  0.8086, Adjusted R-squared:  0.8082
F-statistic:  2547 on 35 and 21110 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 0.2362 on 21122 degrees of freedom
Multiple R-squared:  0.7991, Adjusted R-squared:  0.7988
F-statistic:  3652 on 23 and 21122 DF,  p-value: < 2.2e-16
```

Polynomial outperforms Linear models

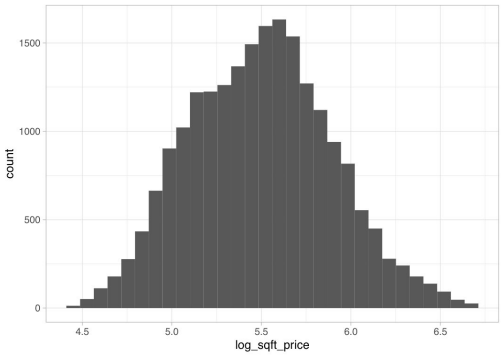
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.827e+02  9.227e+02   0.632 0.527676
date                     5.523e-06  3.674e-07  15.033  < 2e-16 ***
bathrooms                9.135e-02  3.381e-03  27.022  < 2e-16 ***
floors                  -5.352e-02  5.139e-03 -10.416  < 2e-16 ***
waterfront               4.055e-01  2.073e-02  19.561  < 2e-16 ***
view                     7.153e-02  2.567e-03  27.864  < 2e-16 ***
condition                1.418e-01  2.436e-02   5.824 5.85e-09 ***
I(condition^2)          -7.880e-03  3.230e-03  -2.439 0.014725 *
grade                    1.578e-01  2.545e-03  62.001  < 2e-16 ***
yr_built                -1.748e-01  9.405e-03 -18.586  < 2e-16 ***
I(yr_built^2)            4.384e-05  2.403e-06  18.244  < 2e-16 ***
yr_renovated            -5.149e-03  5.217e-04  -9.869  < 2e-16 ***
I(yr_renovated^2)        2.605e-06  2.614e-07   9.966  < 2e-16 ***
zipcode                 -1.157e-03  4.159e-05 -27.818  < 2e-16 ***
long                     1.627e+02  1.489e+01  10.929  < 2e-16 ***
I(long^2)                6.678e-01  6.097e-02  10.953  < 2e-16 ***
lat                      3.990e+02  8.458e+00  47.166  < 2e-16 ***
I(lat^2)                -4.181e+00  8.897e-02 -46.995  < 2e-16 ***
log(sqft_living15)       6.909e-01  1.835e-01   3.764 0.000168 ***
I(log(sqft_living15)^2) -2.767e-02  1.224e-02  -2.262 0.023731 *
log(sqft_lot15)         -2.631e-01  3.325e-02  -7.912 2.65e-15 ***
I(log(sqft_lot15)^2)     1.232e-02  1.766e-03   6.973 3.20e-12 ***
log(sqft_lot)            2.430e-02  5.549e-03   4.379 1.20e-05 ***
I(log(sqft_above)^2)     1.706e-02  5.630e-04  30.305  < 2e-16 ***
```

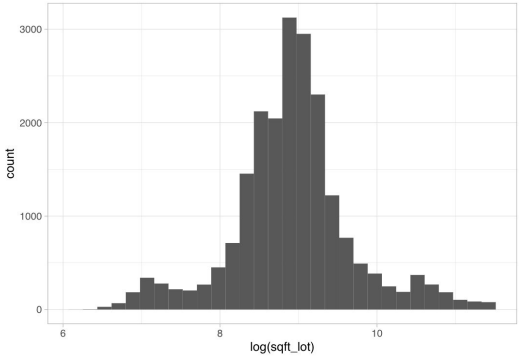
COMPARE PROPORTIONS

Houses with more bedrooms are more likely to have higher sqft_price?

More bedroom: log(bedrooms)>1.099
Higher sqft_price: sqft_price(greater than 5.763)

Yes





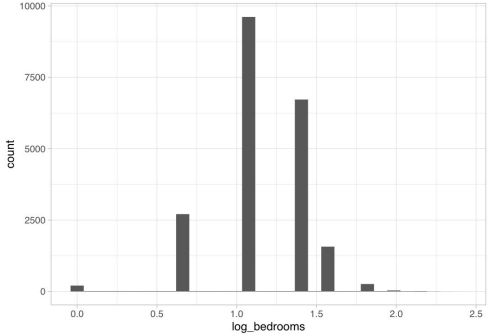
House with greater lot area are more likely to have higher sqft_price?

Greater lot area: log(sqft_lot)>8.924
Higher sqft_price: sqft_price(greater than 5.763)

Yes

	log_sqft_price>5.5	log_sqft_price<=5.5
log_sqft_lot>8.924	1847	8723
log_sqft_lot<=8.924	3494	7082

	log_sqft_price>5.5	log_sqft_price<=5.5
log_bedroom>1.099	1586	7029
log_bedroom<=1.099	3755	8776



PREDICTION

Estimate budget of buying a place in Seattle (county seat)

How about a 2b2b in Bellevue?

```
       fit      lwr      upr
1 602082.7 382151.9 948585.1
```

How's the prediction?

Nah, need updated data:(