Prediction of House Sale Price in King County

Introduction

For decades, house sale price forecasting has always been a hot topic. Unlike macroeconomists who make predictions of the housing market according to the economic situation, people usually would estimate the price of a certain house based on variables such as the size, location, and condition of the house. For my project, I decided to build models on house prices in order to investigate how internal and external factors would influence the sale price. I chose a dataset[1] published on Kaggle in 2016. This dataset includes house sale prices for King County in Washington between May 2014 and May 2015. It's scraped from KingCounty.gov with the dimension of 21 columns and 216,213 rows. Among 21 columns, there only exits one variable(date) that is not numerical, which suggests this data would be a good choice for building regression models based on it.

Findings

1. Location Matters

Location, Location, Location. Most people heard this mantra about the price of properties. In my project, I confirmed this saying by comparing two models' ability of prediction. While model 1 utilizes variables of house size and location, reduced model 2 only uses house size-related variables. As I expected, model 1 shows great improvement than model 2 by explaining 16.75% more about the variation of data (greater R-squared).

[1] https://www.kaggle.com/harlfoxem/housesalesprediction

2. Polynomial Model Wins

Within this project, I built 2 multiple regression models and 2 polynomial models with degrees of 2. After conducting ANOVA test and comparing R-squared values between them, I conclude that polynomial models outperform the multiple regression ones. The best model achieved R-squared value of 0.8082.

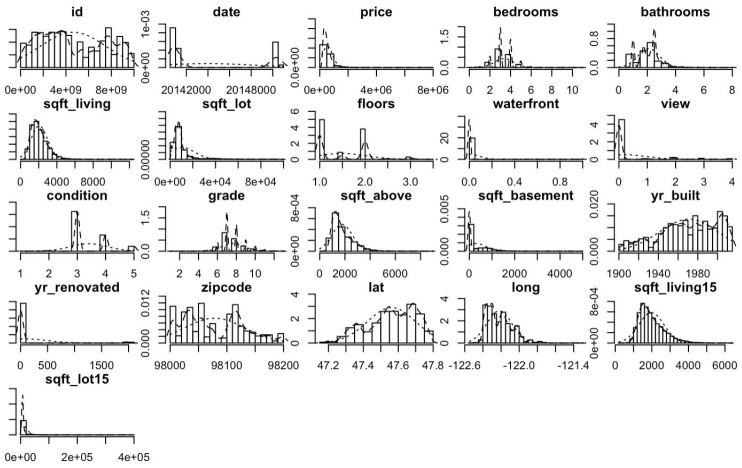
3. Bigger House, More Expensive per Sqft

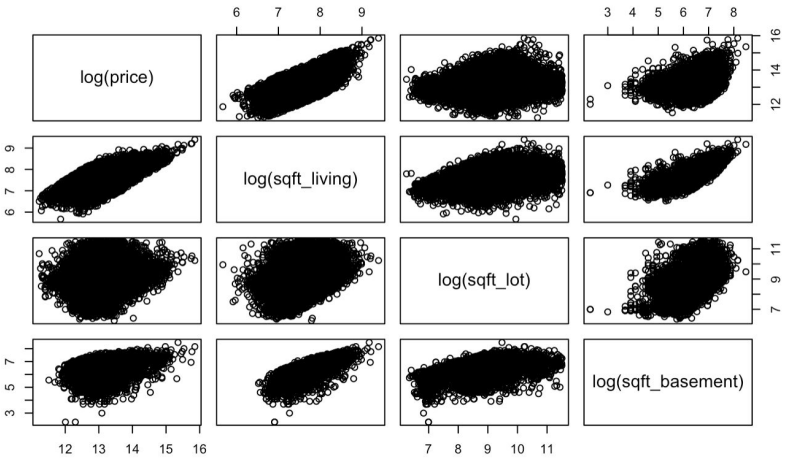
How house price is influenced with the increase in size? Will bigger house be cheaper per sqft? I conducted two proportion comparison of this assumption. Opposite to what I expected, data shows house with greater number of bedrooms or larger lot size both result in higher price per sqft. However, I believe this conclusion might be influenced by the type of property, and testing on apartment sale instead of house sale could give a different answer.

Summary

At first, I created a histogram plot of all variables. We can see some of the variables are close to normal distribution. Furthermore, I did log-transformation for dependent variable price and size-related variables for model building. For model fitting, I built multiple regression models and polynomial models accordingly. Since the polynomial models with the highest R-squared value, we can say that this model successfully explains 80.82% variation of price.

- Histogram of Variables & Log-transformation





- Multiple Regression Model:

```
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
```

- Polynomial Regression Model:

```
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 ***
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

For proportion comparison, I created 2 tables for testing. Both result in p-value of 1 which suggests there is strong evidence that greater number of bedroom (> median of the number of bedroom) and larger lot size (> median of the log(lot_sqft)) will bring expensive price per sqft (> median of the price per sqft).

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

[1] 1

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

[1] 1

In addition, I tried a test case of predicting the price of a 2b2b house built in 2016 in Bellevue, Seattle. The predicted price is $602,082.7 with the 95% confidence interval range from $382,151.9 to $947,585.1. Searched real house price in Bellevue, I found this outcome is not accurate. If we want to forecast current house price, data should be updated since house sale price is quite time-sensitive.

Reference

- Data schema from https://www.slideshare.net/PawanShivhare1/predicting-king-county-house-prices

Variable	Description
Id	Unique ID for each home sold
Date	Date of the home sale
Price	Price of each home sold
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
Sqft_living	Square footage of the apartments interior living space
Sqft_lot	Square footage of the land space
Floors	Number of floors
Waterfront	A dummy variable for whether the apartment was overlooking the waterfront or not
View	An index from 0 to 4 of how good the view of the property was
Condition	An index from 1 to 5 on the condition of the apartment,
Grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
Sqft_above	The square footage of the interior housing space that is above ground level
Sqft_basement	The square footage of the interior housing space that is below ground level
Yr_built	The year the house was initially built
Yr_renovated	The year of the house's last renovation
Zipcode	What zipcode area the house is in
Lat	Lattitude
Long	Longitude
Sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
Sqft_lot15	The square footage of the land lots of the nearest 15 neighbors