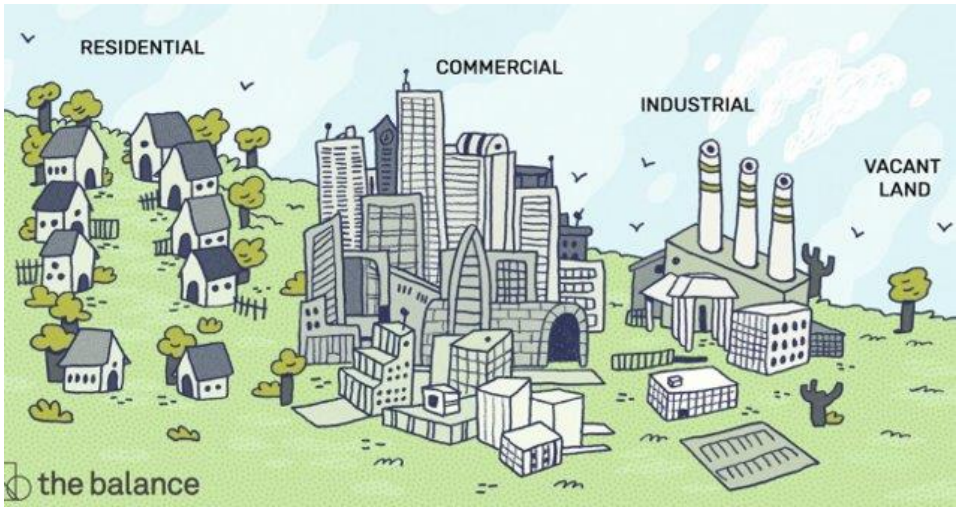

Multiple Linear Regression Analysis: Portland Real Estate Prices

Data 603: Statistical Modeling With Data

University of Calgary

Winter 2021

Group 4 - Sai Lakkavaram, Lu Li, Chang Sun, Bhanu Vedula

Table of Contents

Introduction

Housing price reflects the economy and is of interest to both buyers and sellers. During these times of uncertainty COVID-19 is accelerating suburban – growth and has an impact on overall real estate prices for residential and commercial properties. (Emerging trends in Real Estate® 2021). House price continues to increase year over year.

This report aims to explore variables that most influence price and to create a model to predict price for properties in Portland, USA.

Methodology

Data Source

The dataset for Portland real estate listings was downloaded from www.redfin.com. Redfin permits downloading data from a region of choice if the local MLS allows it. A limit of 350 records is set and can be downloaded in csv format. The dataset does not contain any personal information and is open to use for analysis by third parties. The downloaded data has several quantitative variables like sale price, number of bedrooms, number of bathrooms, listing date, and number of days the listing is active. It also includes qualitative variables like property type and the status of the listing.

According to Redfin, the median sale price of a property in Portland is $520,000 and median days on market is 14. The data we downloaded includes active listings in year 2020 and price ranges from $200K to $600K.

Data Wrangling

Our first attempt at cleaning the data set included deleting unwanted columns for the analysis. These columns include open house data, URL link to images, lot size, price per square feet, HOA/month (Homeowners Association fees), listing status and unique addresses of properties, geographical information like latitude and longitude. The price of the property is chosen as dependent variable. Number of bedrooms, bathrooms, days on the market, size of the property in sq. ft, built year and property type were identified as the independent variables.

The next step was to delete records with NA values. There are 22 records with no information for beds, bathrooms and size. The property type on these records is vacant land. Additionally, there are interesting record that had zero bedrooms. These records were included in the dataset as they are valid listings for studio apartments. There are a few data points with half bathrooms. We chose to keep these records as well since they are commercial listings for spas, salons and warehouses.

Variable Explanations

The following is a complete list of variables used in our modelling:

1. PRICE– Price of the property (*in dollars)* *Dependent variable
2. PROPERTY TYPE – Type of the property (Condo, Single Family, Multi-family, etc.) *Categorical Independent Variable

3. BEDS – Number of bedrooms *Quantitative Independent Variable
4. BATHS- Number of bathrooms * Quantitative Independent Variable
5. SQUARE FEET – Size of the property (in square feet) * Quantitative Independent Variable
6. YEAR BUILT- Year the property was built in (Year) * Quantitative Independent Variable
7. DAYS ON MARKET – Number of the days the listing was on MLS * Quantitative Independent Variable

Data Modelling

The modelling approach of this project is built on the multiple linear regression modelling techniques we learned in Data 603. The dataset was split into training (80%) and testing datasets (20%). After identifying interested independent variables, they were tested for multi-collinearity. We built the ordinary least squares linear model with all the predictors we were interested in and identify the best variables using automatic selection methods.

We tuned model with interactions terms and higher order terms. We did an ANOVA test for the additive model and high order model. Our final model with higher order terms was then tested for the following five assumptions as shown below:

1. Linearity Assumption - Review residual plots
2. Normality Assumption - Using Shapiro-Wilk normality test
3. Equal Variance Assumption (heteroscedasticity) - Using Breusch-Pagan test
4. Multi-collinearity - Using variance inflation factors (VIF)
5. Outliers - check Cook's distance and leverage

The final model that satisfies all the assumptions was then used to predict house prices.

Results

Identifying main effects

The first-order model was built using all the variables in the dataset.

$$Y_{Price} = \beta_0 + \beta_1 X_{PropertyType} + \beta_2 X_{Beds} + \beta_3 X_{Baths} + \beta_4 X_{SquareFeet} + \beta_5 X_{YearBuilt} + \beta_6 X_{DaysOn\ Market}$$

This model is then used to identify the significant variables further along using forward regression, backward regression, stepwise regression and all possible selection procedures. Forward regression, backward regression, stepwise regression all selected Beds, Baths and Square. Ft as significant variables. Summary is detailed as below:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 247697.63   13343.89  18.563  < 2e-16 ***
BEDS        -16165.69    7278.22  -2.221   0.0272 *
BATHS        36129.16    9132.64   3.956 9.87e-05 ***
SQUARE.FEET    108.98      13.28   8.209 1.10e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77820 on 256 degrees of freedom
Multiple R-squared:  0.4399,    Adjusted R-squared:  0.4334
F-statistic: 67.03 on 3 and 256 DF,  p-value: < 2.2e-16
```

Figure 1: Regression Result

All possible regression method gives the best selection for different number of variables combination:

```
            Cp      AIC     adjr2
[1,] 18.984155 6612.025 0.4023840
[2,]  9.690452 6603.121 0.4246984
[3,]  5.021761 6606.461 0.4281217
[4,]  1.224206 6602.529 0.4387944
[5,]  1.468751 6602.706 0.4404867
[6,]  3.000000 6604.217 0.4392952
[1] "SQUARE.FEET"
[2] "BATHS SQUARE.FEET"
[3] "PROPERTY.TYPE BATHS SQUARE.FEET"
[4] "PROPERTY.TYPE BEDS BATHS SQUARE.FEET"
[5] "PROPERTY.TYPE BEDS BATHS SQUARE.FEET YEAR.BUILT"
[6] "PROPERTY.TYPE BEDS BATHS SQUARE.FEET YEAR.BUILT DAYS.ON.MARKET"
```

Figure 2: All possible Best Selection

Selection with 4, 5 and 6 variables all have Mallow's Cp value less than their number of variables and can be considered as good candidates. Among these candidates, selection with 4 variables has the lowest AIC, meaning it can fit the data with least complexity. Therefore, it's chosen as the best candidate. Since it has an extra variable, Property Type, compared with the selection made from previous three methods, an ANOVA test is used to compare these two selections. ANOVA test result is shown as below:

```
Analysis of Variance Table

Model 1: PRICE ~ BEDS + BATHS + SQUARE.FEET
Model 2: PRICE ~ PROPERTY.TYPE + BEDS + BATHS + SQUARE.FEET
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1    256 1.5505e+12
2    251 1.5056e+12  5 4.4833e+10 1.4948  0.192
```

Figure 3: ANOVA Result

P value is 0.192, greater than 0.05, indicating this extra variable is not significant. Therefore, it's not included in the model.

Considering the fact that houses stay longer on market tend to sale at a lower price. Variable Days on Market should be included as a variable. In the end final additive model is:

$$Y_{Price} = \beta_0 + \beta_1 X_{Beds} + \beta_2 X_{Baths} + \beta_3 X_{SquareFeet} + \beta_4 X_{DaysOn\ Market}$$

Interaction Terms Individual T-tests

To tune the model, we want to find if there is any interaction between variables. A hypothesis test on interaction terms was conducted.

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

*where i = Beds*Baths, Beds*SquareFeet, Beds*DaysOnMarket, Baths*SquareFeet, Baths* DaysOnMarket or SquareFeet *DaysOnMarket*

Result is as below:

*Beds*Baths: p* = 0.198
*Beds*SquareFeet: p =0.112*
*Beds*DaysOnMarket: p =0.849*
*Baths*SquareFeet: p=0.319*
Baths DaysOnMarket: p=0.935*
*SquareFeet *DaysOnMarket: p=0.852*

Individual T-tests were used to determine the best predictors based on a significance level of $\alpha = 0.05$. From the results of these tests, at 95 % significance level we fail to reject the null hypothesis. This suggests that there is no interaction between variables.

Higher Order Terms Individual T-tests

To further tune the model, we want to check for higher order terms.

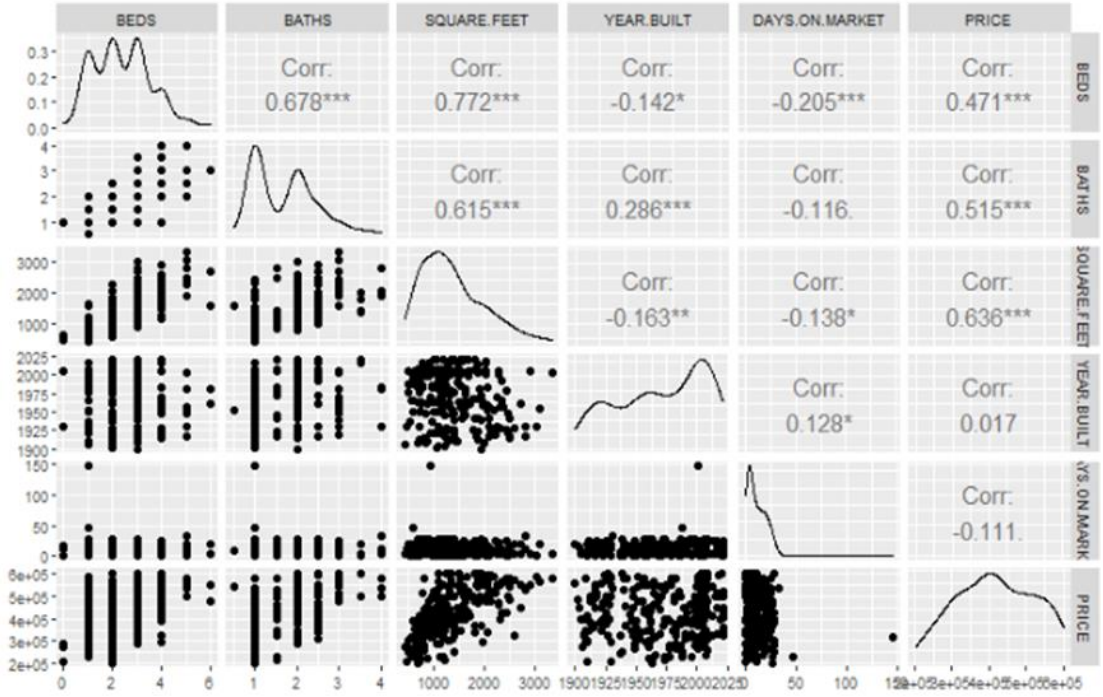

Figure 4: Scatter Plot Matrix

From the scatter plot matrix, it can be seen there is a downward curvature relation between Price and Square. Feet. A quadratic term of Square. Feet is added to account for this relation. A hypothesis test on this quadratic term was conducted.

$$H_0: \beta_{Square.Feet\text{^}2} = 0$$
$$H_a: \beta_{Square.Feet\text{^}2} \neq 0$$

P value is 6.19e-05, way lower than 0.05, indicating this quadratic term is highly significant. To avoid missing out other curvature relations, individual t-test was conducted on all possible quadratic terms:

$$H0: \beta_i = 0$$
$$Ha: \beta_i \neq 0$$

where i = Beds^2, SquareFeet^2, DaysOnMarket^2 or Baths^2

Result is as below:

Beds^2: p = 0.06514
SquareFeet^2: p = 6.48e-06
DaysOnMarket^2: p = 0.55687
Baths^2: p= 0.66064

Other than Square. Feet, quadratic terms of other variables are not significant.

Hypothesis Statement for ANOVA Test

$$H_0: \beta_{Square.Feet\text{^}2} = 0$$

$$H_a: \beta_{Square.Feet\text{^}2} \neq 0$$

An ANOVA test compared our reduced model (main effect) with the full model (main effect and higher order). From the results of the ANOVA (F = 16.594, p = 6.192e-05), we reject the null hypothesis. Also adjusted R square of the additive model is 0.4323. Adding a Square.Feet^2 term increases adjusted R square to 0.465. We conclude that the higher order term for square feet is significant and should be included in our final model. Therefore, our model is:

$$Y_{Price} = 1.554e{+}05 + 2.736e{+}02\, X_{SquareFeet} - 2.169e{+}04\, X_{Beds} + 3.013e{+}04\, X_{Baths} - 2.359e{+}02\, X_{DaysOnMarket} - 4.791e{-}02\, X_{SquareFeet}^2$$

Multiple Regression Assumptions

Whether or not this model can be used to explain the data and prediction price relies on whether it meets linear regression assumption. We testify all assumptions as below.

Linearity Assumption and Independent Assumption

In this Residual vs. Fitted Values plot, residuals scatter above and below the line residual=0 randomly. We don't see any specific pattern in the residuals. Also, one residual doesn't affect another, suggesting that linearity assumption and independent assumption are met.

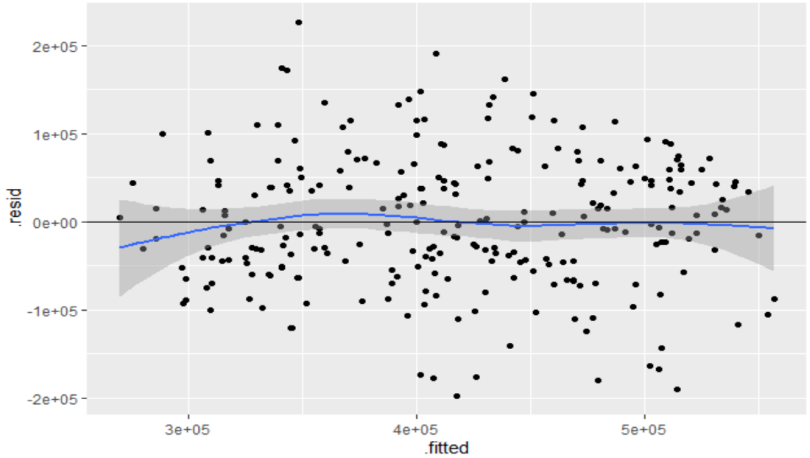

Figure 5: Residual vs Fitted Values Plot

Normality Assumption

One of the assumptions of linear regression is that the residuals are normally distributed.

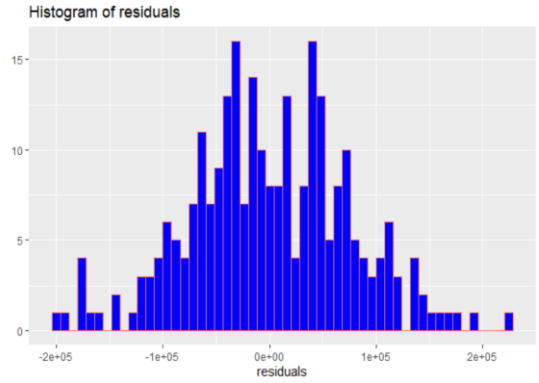

Figure 6: Histogram of Residuals

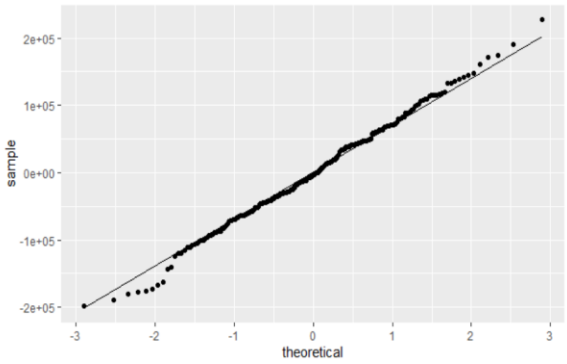

Figure 7: Q-Q Norm Plot

Histogram of residuals shows that it follows a normal distribution roughly. From QQ-plot, residuals line up with the normal distribution line pretty good, except for a few points at bottom. To confirm if these points violate normal distribution, Shapiro-Wilk test is used:

H_0: The residuals are normally distributed
H_a: The residuals are not normally distributed

```r
#Testing for Normality
shapiro.test(residuals(model_order))
```

```
        Shapiro-Wilk normality test

data:  residuals(model_order)
W = 0.99609, p-value = 0.7651
```

Figure 8: Shapiro-Wilk Test Result

P value of Shapiro-Wilks test is 0.7651. Since the p-value is greater than 0.05, we fail to reject the null hypothesis. Residuals are normally distributed.

Equal Variance Assumption

Another assumption is residuals should have the same variance.

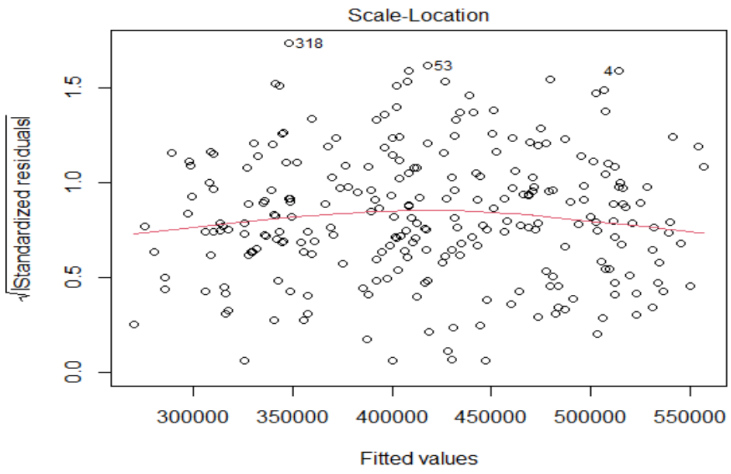

Figure 9: Scale-Location Plot

Trend line in Scale-location plot is almost a horizontal line, meaning these is no change of variance when fitted values changed. We confirmed this with Breusch-Pagan test:

H_0: Heteroscedascity is not present
H_a: Heteroscedascity is present

```
```{r}
library(lmtest)
bptest(model_order)
```
```

```
        studentized Breusch-Pagan test

data:  model_order
BP = 8.6056, df = 5, p-value = 0.1259
```

Figure 10: Breusch-Pagan Test Result

P value of Breusch-Pagan test is 0.1259. Based on α=0.05, we fail to reject the null hypothesis. Equal variance assumption is met.

Multi-collinearity Tests

To test for multi-collinearity in our model, we looked at multiple variance inflation factors (VIF) to determine which independent variable has high collinearity relation with others. VIF result is below:

```
 VIF Multicollinearity Diagnostics

                     VIF detection
SQUARE.FEET      27.7933         1
BEDS              3.1164         0
BATHS             1.9811         0
DAYS.ON.MARKET    1.0475         0
I(SQUARE.FEET^2) 22.7519         1

Multicollinearity may be due to SQUARE.FEET I(SQUARE.FEET^2) regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Figure 11: VIF for Model_order

SQUARE.FEET and I (SQUARE.FEET^2) have greater-than-10 VIF values. It makes sense because one is highly correlated with its quadratic term. However, this doesn't necessarily mean the model has multi-collinearity issue. We checked the VIF values on main effects again and result is below:

```
 VIF Multicollinearity Diagnostics

                   VIF detection
SQUARE.FEET    2.5802         0
BEDS           3.0377         0
BATHS          1.9251         0
DAYS.ON.MARKET 1.0458         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

Figure 12: VIF for Main Effects

None of predictors have VIF value greater than 5, indicating our model doesn't have multi-collinearity issue.

Influential Points and Outliers

Influential points can greatly affect the coefficients of our model. To check for influential points, standardized residuals vs leverage is plotted:

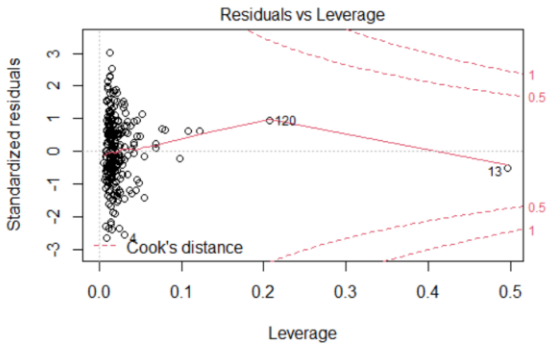

Figure 103: Standardized Residuals vs. Leverage

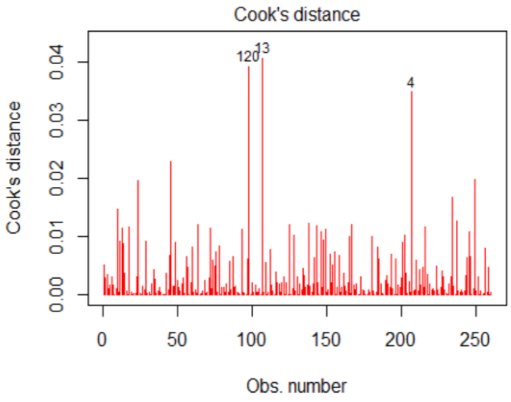

Figure 11: Cook's Distance Plot

It can be seen there is no data points beyond the dash line of Cook's distance, indicating there is not any influential points.

Although there aren't any influential points, there are points with high leverage. If points with leverage greater than 2p/n are considered as high leverage points, where p is number of coefficients and n is number of data points, there are 15 high-leverage points. If points with leverage greater than 3p/n are considered as high leverage points, there are 7 high-leverage points. This is illustrated in below leverage plot and output:

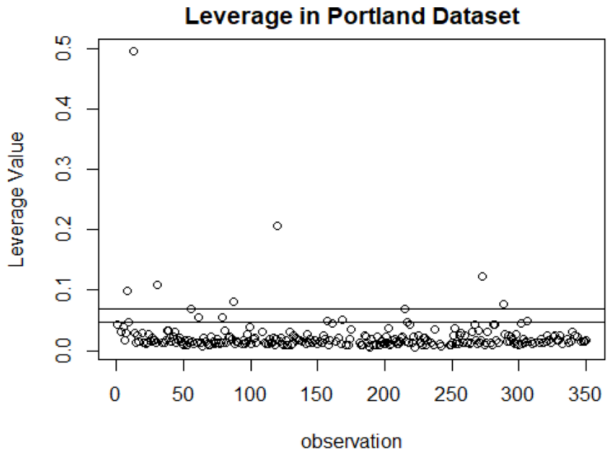

Figure 125: Leverage Points Plot

```
[1] "** 2p/n Outliers **"
       168        157        120         13        217         31        288        273         61         56        215          8
0.05173571 0.04925427 0.20743219 0.49620811 0.04631913 0.10836473 0.07616801 0.12193862 0.05463608 0.06890874 0.06869136 0.09802471
        87         79        306
0.08076417 0.05516674 0.04908000
[1] "                  "
[1] "                  "
[1] "** 3p/n Outliers **"
       120         13         31        288        273          8         87
0.20743219 0.49620811 0.10836473 0.07616801 0.12193862 0.09802471 0.08076417
```

We tried removing these high-leverage points and wanted to see if this would improve model performance. After removing points with leverage greater than 2p/n, adjusted R^2 is 0.4388. After removing points with leverage greater than 3p/n, adjusted R^2 is 0.4488. Compared with 0.465 of original model, they are reduced quite a bit. Therefore, these high leverage points are kept considering model performance.

```
Residual standard error: 76710 on 239 degrees of freedom
Multiple R-squared:  0.4503,    Adjusted R-squared:  0.4388
F-statistic: 39.16 on 5 and 239 DF,  p-value: < 2.2e-16
```

*Figure 136: Model Performance After Removing Points With Leverage >2*p/n*

```
Residual standard error: 76060 on 247 degrees of freedom
Multiple R-squared:  0.4597,    Adjusted R-squared:  0.4488
F-statistic: 42.03 on 5 and 247 DF,  p-value: < 2.2e-16
```

*Figure 146: Model Performance After Removing Points With Leverage >3*p/n*

Interpreting Coefficients

Beta1 = 2.736e+02: Since there is quadratic term of Square. Feet in the model. Beta 1 should not be interpreted.

Beta2 = -2.169e+04: If a house increases bedroom by 1, price decreases by 2.169e+04 dollars, holding other independent variables unchanged.

Beta3 = 3.013e+04: If a house has 1 more bathroom, price increases by 3.013e+04 dollars, holding other independent variables unchanged.

Beta4 = -2.359e+02: If days on market of the house increase 1 day, price drops 2.359e+02 dollars, holding other independent variables unchanged.

Beta5 = -4.791e-02: Negative coefficient indicates there is a downward curvature relation between Price and Square. Feet.

Adjusted R^2=0.465: 46.5% of price variable can be explained by number of bedrooms, bathroom, square feet and days on market.

Predicted House Prices

To check if our model is good for predicting the house price, we did a prediction on test data. Since data has multiple variables, it can't be visualized by multi-dimensional graph. Prediction price is plotted against each independent variable as below:

The Actual vs Predicted plots. Each variable has its own plot.

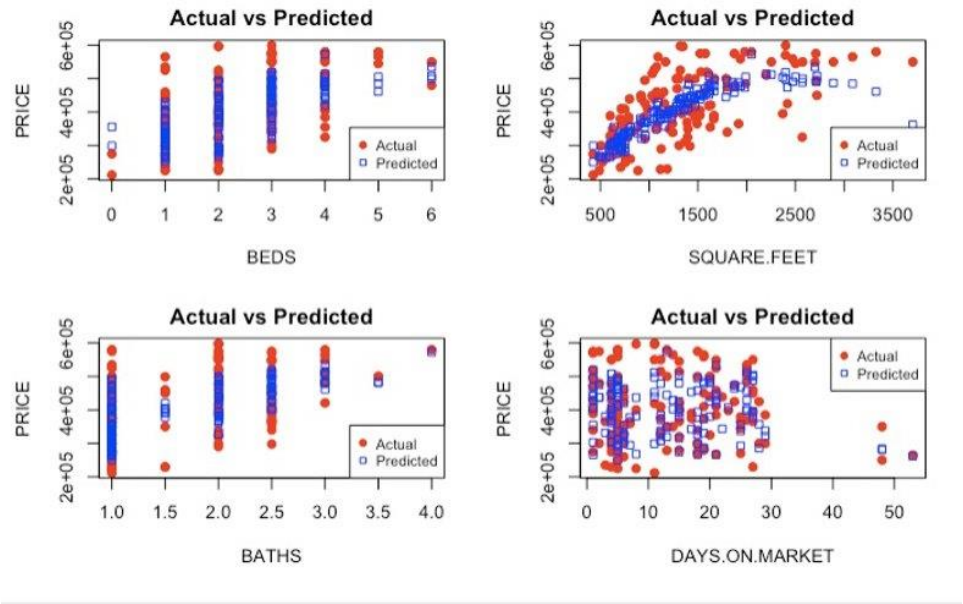

Figure 157: Actual vs Predicted Plot For Each Variable

From the Actual data vs Predicted data plot, we could see that predicted data overlie actual data in general.

We also draw the 95% confidence prediction intervals, again each variable has its own plot.

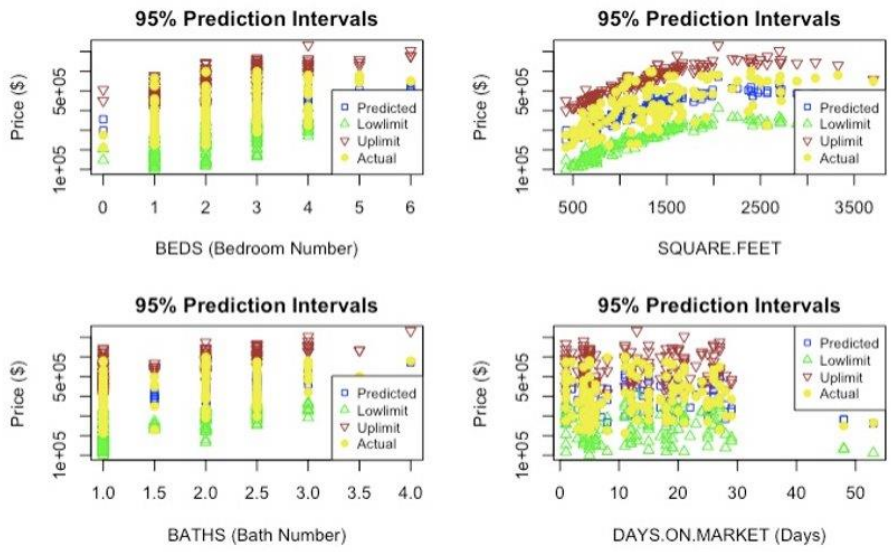

Figure 18: 95% Confidence Interval For Each Variable

From the 95% prediction intervals plots, we could see that all the predicted data are between the upper limit and lower limit.

The Predicted Price is plotted against Actual Price plot:

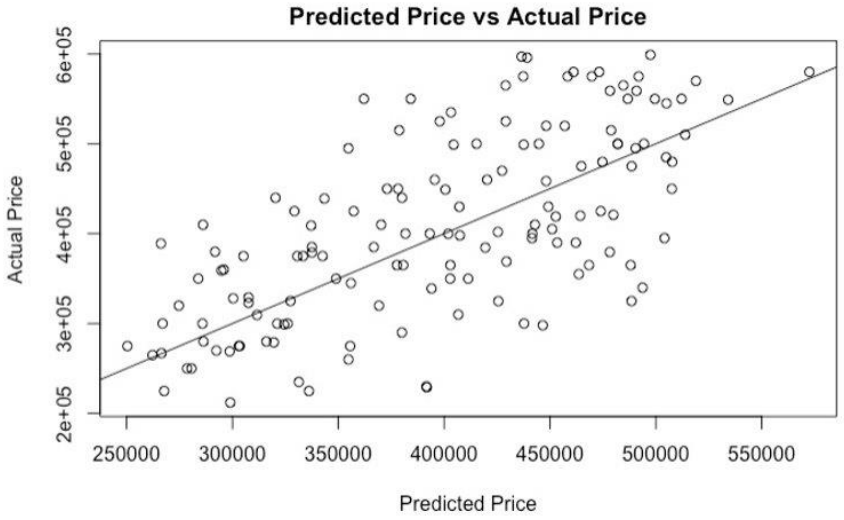

Figure 19: Predicted Vs Actual Price Plot

It can be seen that all the points are around the line of Actual Price=Predicted Price.

Calculated Mean Absolute P Percentage Error value is 0.1582974.

Conclusion

With limited independent variables available, model best to describe and predict Portland real estate price is:

$$Y_{Price} = 1.554e{+}05 + 2.736e{+}02\ X_{SquareFeet} - 2.169e{+}04\ X_{Beds} + 3.013e{+}04\ X_{Baths} - 2.359e{+}02\ X_{DaysOnMarket} - 4.791e{-}02\ X_{SquareFeet}^2$$

Discussion

While real estate pricing largely depends on a multitude of factors, the factors we chose provided reasonable results for our analysis.

One of the challenges we encountered while working on this project was finding a dataset. In reality, real estate price studies result in large price variations with large ranges and tend to lead to heteroscedasticity, violating the equal variance assumptions.

Inclusion of additional meaningful factor like neighborhood characteristics, distance to amenities, school and crime rate would have resulted in different and probably more useful model.

Although the usability of the model depends on its intended use, there are several ways we could improve our model. As mentioned above, including additional variables that influence the price of a property and utilizing a more advanced statistical analysis methods like weighted average regression would definitely improve the true explanatory power of the model.

References

- Hardin, J. Math 58B - Introduction to Biostatistics. Example R code / analysis for housing data. http://pages.pomona.edu/~jsh04747/courses/math58/Final_examp.html.
- Li, H., & Gan, S. N. MATH1312 Regression Analysis Project. RPubs. https://www.rpubs.com/sngan/517214.
- Shiny. (2016). Predicting House Prices. rstudio-pubs-static.s3.amazonaws.com. https://rstudio-pubs-static.s3.amazonaws.com/150743_fbe2be64165349798440e35351653b16.html.
- Smalley, H. K. (2019, September 9). Class 3B - One Sample Bootstrap. RPubs. https://rpubs.com/hsmalley/math239-3b.
- Timbers, T.-A., Campbell, T., & Lee, M. (2021, January 12). Data Science: A First Introduction. Chapter 9 Regression II: linear regression. https://ubc-dsci.github.io/introduction-to-datascience/regression2.html#overview-7.
- Utilizing Linear Regression to Estimate Real Estate Prices. rstudio-pubs-static.s3.amazonaws.com. (2018, December). https://rstudio-pubs-static.s3.amazonaws.com/448025_2ee903e41949413ea65ff19f1e6700d7.html.
- Yuan, L. (2019). A REGRESSION MODEL OF SINGLE HOUSE PRICE IN LA CONSTRUCTING A PREDICTED MODEL FOR HOUSE PRICES. https://scholarworks.calstate.edu. https://scholarworks.calstate.edu/downloads/dj52w646n.
- Simovic, D. (2021, January 6). 30+ Essential Real Estate Statistics - 2021 Edition. SmallBizGenius. https://www.smallbizgenius.net/by-the-numbers/real-estate-statistics/#gref
- Downloadable housing market data - redfin. (2020, October 16). Retrieved April 01, 2021, from https://www.redfin.com/news/data-center/