King County House Price

DSC 324, Advance Data Analysis

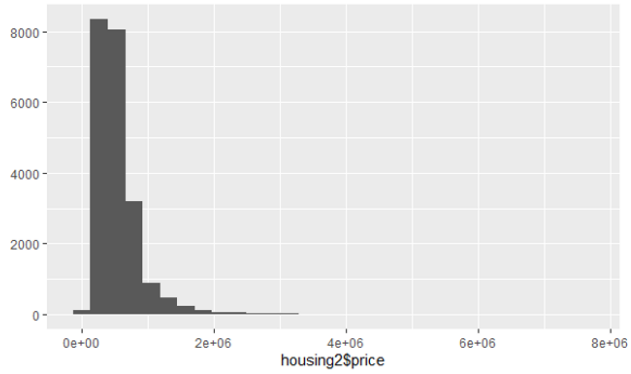
Saransh Thakur, Kavit Patel, Jessenia Jimenez

Introduction

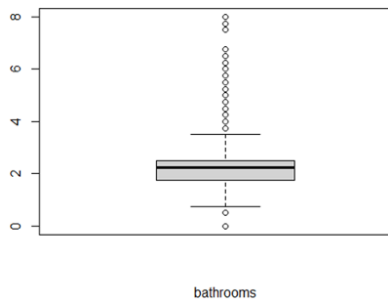
The goal of this report is to identify the relationship between house price in King County and the condition of the house given several factors including: number of bedrooms, number of bathrooms, square footage of living room, square footage of the lot, total floors in house, square footage of house apart from the basement, square footage of house including basement, year built, year when the house was renovated, living room area in 2015, and lot area in 2015. We will also test whether certain conditions such as whether the house has a waterfront and whether there is an outside view affect the price of a house.

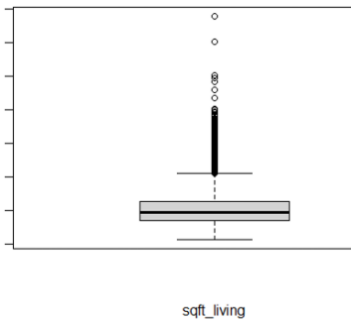
Exploratory Analysis

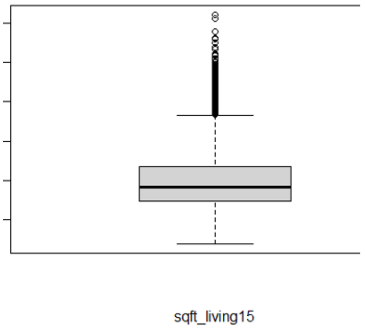
When creating a histogram for the distribution of each of the predictors and price, we noticed distribution issues.

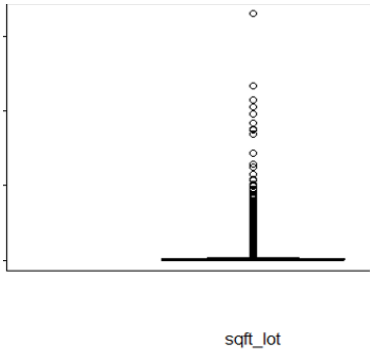


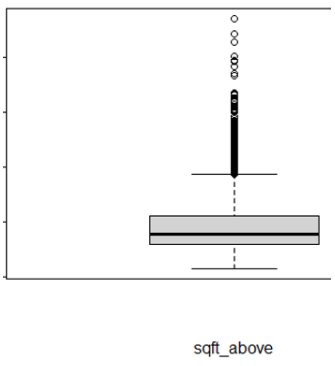
In the histogram for price, there is a positive skew to the right, indicating that most of the data is less than the mean. To combat, and normalize the data, we transformed price using log transformation for a better-fitted model and stabilized the variance.

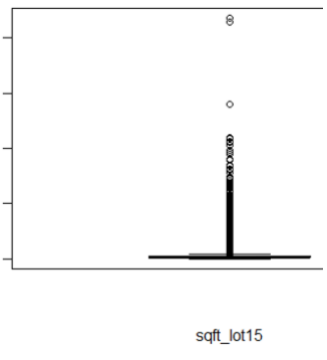












We noticed that many of the predictors were also skewed which could influence a bad model. To combat and normalize the data, we decided to also transform the predictors above using log transformation (*bedrooms, sqft_living, sqft_living15, sqft_lot, sqft_above, and sqft_lot15*). This can clean out the issues that could occur when applying different techniques to the dataset.

Application of Analysis

The analysis of the dependent variable analysis is divided into three performances: regularized regression, principal component analysis, and linear discriminant analysis. Using both regularized regression and principal component analysis to calculate numeric performances on the dependent variable, price, and linear discriminant analysis to calculate performance on the categorical variable, condition.

Regularized Regression

This section will analyze the strength of prediction in price by evaluating the application of Ordinary Least Squares Regression and several regularized regression techniques such as Ridge Regression, Lasso Regression, and Elastic Net Regression. We start by randomly separating 70% of the data into a training set and 30% into a test set. In doing so, we can differentiate and compare the resulting RMSE and R-squared performance values of each set.

OLS REGRESSION

After transforming the dependent variable price (log_price), bathroom, square-foot of living room, square-foot of the house apart from the basement, square-foot of the living room in 2015, and square-foot of the lot in 2015. We ran OLS regression modeling to predict log_price and the following is the resulting equation:

```
Call:
lm(formula = log_price ~ ., data = housingTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36033 -0.20152  0.01291  0.20368  1.30118

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       18.4668091  0.2387923  77.334  < 2e-16 ***
bedrooms          -0.0298336  0.0034900  -8.548  < 2e-16 ***
floors             0.0749151  0.0071462  10.483  < 2e-16 ***
waterfront         0.4178676  0.0320619  13.033  < 2e-16 ***
view               0.0457365  0.0037707  12.129  < 2e-16 ***
condition          0.0394924  0.0042181   9.363  < 2e-16 ***
grade              0.2136384  0.0036937  57.839  < 2e-16 ***
yr_built          -0.0056691  0.0001156 -49.040  < 2e-16 ***
log_bathroom       0.1797793  0.0182407   9.856  < 2e-16 ***
log_sqft_living    0.3995146  0.0165224  24.180  < 2e-16 ***
log_sqft_lot      -0.0100514  0.0072972  -1.377    0.168
log_sqft_above    -0.0943997  0.0149980  -6.294 3.18e-10 ***
log_sqft_living15  0.2642779  0.0126843  20.835  < 2e-16 ***
log_sqft_lot15    -0.0516465  0.0079777  -6.474 9.85e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3072 on 15115 degrees of freedom
Multiple R-squared:  0.6568,    Adjusted R-squared:  0.6565
F-statistic:  2225 on 13 and 15115 DF,  p-value: < 2.2e-16
```

log_price= -.0298bedrooms + .0749floors + .4179waterfront + .0457view +.0395condition +.2136grade - .0057 yr_built + .1798 log_bathroom + .3995log_sqft_living - .0100log_sqft_lot - .0944log_sqft_above + .2643log_sqft_living15 -.0517log_sqft_lot15

The results of the OLS Regression indicate the significance of each predictor with log_price, the R-squared value of the model, and the Residual Standard Error. To evaluate the significance of each predictor, we must transform the variables back to its original measures by finding the exponent, subtract it by 1 and multiply by 100. By doing so, we can determine that *waterfront, log_sqft_living, log_sqft_living15, grade, log_bathroom, floors*, *view,* and*, condition* have a positive relationship with log_price. We can also determine that *yr_built, log_sqft_lot, bedrooms, log_sqft_lot15, log_sqft_above* have a negative relationship with log_price. Both in accending order of correlation. Based on the findings, we can determine that if the house has a waterfront, it will affect the price of a house by $51.87 which has the greatest affect overall.

REGULARIZED REGRESSION

The OLS regression did not provide an excellent fitted model as the R-squared value was low at .6568 and there was an insignificant predictor (log_sqft_lot) as its p-value (.168) was higher than alpha (.05). We decided to run regularized regression to determine if there are other underlying issues and to determine whether there are more predictors that are not best for the model. After running the ridge regression, lasso regression, and elastic net regression models, we plot the lambda value against the train data set. We chose a lambda range from 0 to 5 with an interval of .5 increase for ridge and lasso regression while changing the interval to .3 for elastic net regression. At the same time, using alpha values of 0,1, and .05 respectively. Each plot indicates the optimal lambda value for the best fit. The following plots are the Mean-Squared Error-values against Log(λ) values:

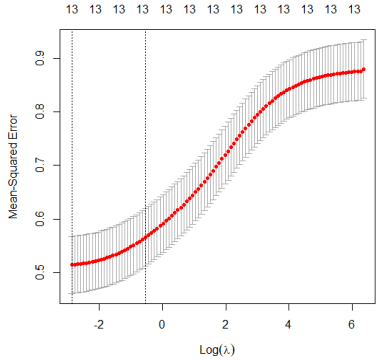


Figure 1-Ridge Regression

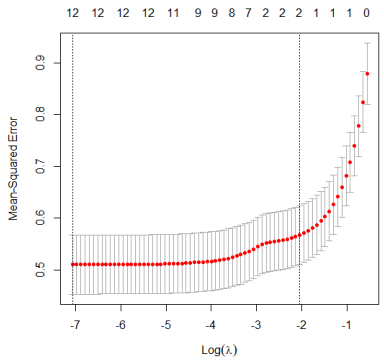


Figure 2-Lasso Regression

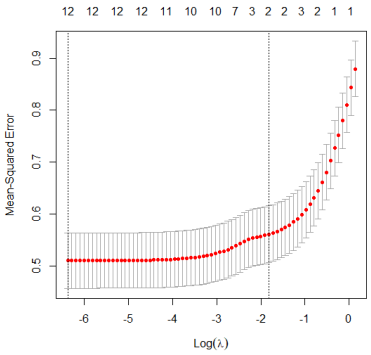


Figure 3-Elastic Net Regression

Method	**log_price R-squared**	**log_price RMSE**
OLS Regression	.6568	0.3073656
Ridge Regression	.4175	0.6670638
Lasso Regression	.4184	0.7080583
Elastic Net Regression	.4210	0.6628633

When evaluating each regression method, we calculate the R-squared value and RMSE value to identify which regression model best fits for log_price. OLS produced the highest r-squared value of .6568, indicating that 65% of the variation in log_price can be explained by the

regression line of log_price on the independent variables. OLS Regression also produced a low RMSE value of .3074 indicating that the predicted and observed data are close to each other. For the other regression analysis, Ridge Regression created the lowest R-squared value, and Lasso Regression created the highest RMSE values. OLS Regression produced a better model with higher accuracy.

Principal Factor Analysis

The PCA components are used as factors in principal factor analysis. In PFA, the loadings are calculated statistically. Here, we maintain only a few components because the others are unexplainable noise and we can examine which component we want using a scree plot.
In PFA, we frequently accept less collected variance (60-80% is generally sufficient).

ROTATION THE COMPONENT USING VARIMAX

We're rotating the component using varimax to reduce minor loadings while enhancing large ones. The divisions are made as clear as feasible by rotating them. Rotation is one method for dealing with the complexity of real-world data, and it greatly enhances interpretability.

```
> p = principal(ds,rotate = "varimax", nfactors=4)
> print(p$loadings, cutoff=.4,sort = T)

Loadings:
                RC1    RC2    RC4    RC3
bedrooms         0.686
grade            0.805
log_bathroom     0.817
log_sqft_living  0.918
log_price        0.766
log_sqft_above   0.853
log_sqft_living15 0.756
log_sqft_lot            0.948
log_sqft_lot15          0.950
floors           0.491          0.518
condition                      -0.802
yr_built                        0.752
waterfront                             0.818
view                                   0.793

                 RC1   RC2   RC4   RC3
SS loadings    4.982 2.139 1.775 1.522
Proportion Var 0.356 0.153 0.127 0.109
Cumulative Var 0.356 0.509 0.635 0.744
>
```

The rotatable components with variable loadings are depicted in the figure above.
We did this since the cut-off of 0.4 makes division plain and easy to grasp. The components are sorted by their eigenvalues, with accumulated variance captured in each component.

The first component (RC1) is mostly composed of positive contributions from the bedroom, grade, bathroom, and sq ft living, sqft above, and sq ft living 15, floor. This component recognizes that the property's price has risen because of this component.

The second component (RC2) has two positive contributions, sqft lot and sqft lot 15. This might imply that a larger sqft lot leads to a higher price, and as a result, houses are more expensive.

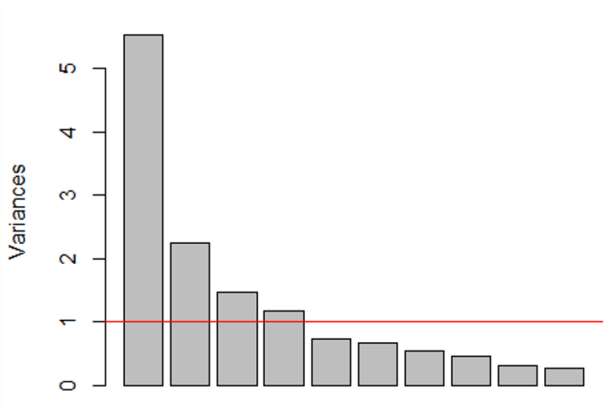
The third component (RC4) is made up of two positive contributions, floor and yr built, and one negative contribution, condition. This may imply a collection of houses with lower pricing since they are in poor shape or were built late with only one story.
The last component (RC3) includes all beneficial impacts from the waterfront and view.

Scree Plot: -

When the scree plot seems to “bend at the knee”, that’s about how many PCs do we want in latent factor discovery; the long tail in the plot tells us about the noise instead of some useful underlying trend.

OR

We can use Principal Components to have an average variance greater than equal to 1.



We utilized an average variance greater than one in the above plot, and we can see from the above scree plot that the first four factors are above the red line, suggesting that we can extract four components.

Linear Discriminant Analysis

After transforming the variables and based on the new dataset, we performed LDA to distinguish and classify the categorial variable in the dataset. There are several variables like waterfront, view, condition, grade, and yr_built that we could use to perform LDA. We performed Linear Discriminant Analysis using the predictor, condition.

Firstly, we removed the log_price variable because the focus is to look at the condition variable and interpret on the dataset, not how condition affects the price. Since the focus of this project is to see how all predictors in the dataset affect price. We took one of the predictors and started evaluating it with other independent predictors. We decided to chose condition as we believe the condition of the house can have a greater affect rather than view, grade, or quality of the building goes from 1-12 which is going to be harder to interpret at this moment.

There are 2 tools that we can use, the Confusion matrix and ROC Curve to interpret. The ideal LDA includes both a confusion matrix and ROC Curve. The ROC curve works on binary classification. If not, there could be issues where the library used works on Binary classification only. This means it only works on 0's and 1's but condition has 1-5. The goal here for LDA is to maximize the distance between 2 categories while minimizing the scatter. Below is more better look at the predictor.

```
> table(new_cleandata$condition)

    1     2     3     4     5 

   30   172 14031  5679  1701 
```

Then we split the data into Training data and Testing Data to perform. Firstly, we performed LDA on condition using the Training Data, then we performed LDA on condition using the Testing Data.

```
> newdataTrain.lda = lda(condition ~ .-log_price, data=newdataTrain)
> newdataTrain.lda
Call:
lda(condition ~ . - log_price, data = newdataTrain)

Prior probabilities of groups:
         1          2          3          4          5
0.001503759 0.008155003 0.649681897 0.262058994 0.078600347

Group means:
  bedrooms   floors  waterfront       view    grade yr_built log_bathroom log_sqft_living
1 2.384615 1.134615 0.038461538 0.34615385 5.769231 1933.769    0.7496154        6.972308
2 2.879433 1.163121 0.007092199 0.07801418 6.489362 1947.993    0.8672340        7.175106
3 3.366865 1.620716 0.006142616 0.21249889 7.833259 1979.411    1.1373373        7.582884
4 3.358640 1.258552 0.009048775 0.26484220 7.391966 1958.361    1.0376981        7.496992
5 3.473878 1.272627 0.013980868 0.32965416 7.318617 1946.700    1.0770272        7.527873
  log_sqft_lot log_sqft_above log_sqft_living15 log_sqft_lot15
1     9.417692       6.916154          7.438846       9.034231
2     9.395816       7.070567          7.342270       9.167660
3     8.910906       7.457052          7.569161       8.890570
4     9.180689       7.299141          7.504211       9.135359
5     8.986946       7.268712          7.467550       8.958955

Coefficients of linear discriminants:
                         LD1         LD2         LD3           LD4
bedrooms          -0.11590005  0.04597764  0.20967961  0.1657085050
floors             0.54688549  0.02887097 -1.47281499  0.3845400661
waterfront        -0.28511481 -0.11516446 -1.40559368 -1.6619298098
view              -0.02536670 -0.02297072 -0.01200565 -0.3524046732
grade              0.11128902  0.26417479  0.56717452  0.3274284090
yr_built           0.03012645 -0.01630044  0.01725388  0.0009629794
log_bathroom      -0.92650025  2.56550301 -2.22439247  0.6546941284
log_sqft_living   -0.56293450  1.58165561  1.10912808 -0.1159883182
log_sqft_lot       0.03396546 -1.09136700 -1.09722982  0.5192286546
log_sqft_above     0.59266073 -0.65041240  0.13499912  0.1904792395
log_sqft_living15  0.29059427 -1.41688323 -1.16422914 -3.9965818878
log_sqft_lot15    -0.33046012  0.61697236  1.32928039  0.0216539295

Proportion of trace:
   LD1    LD2    LD3    LD4
0.8585 0.0954 0.0415 0.0046
```

This tells that LD1 accounts for 84% of the trace or separation and it gives the equation. LD1 is accounting for the most variation in the categories.

*LD1 = - 0.12*bedrooms + 0.55*floors - 0.29*waterfront – 0.03*view + 0.11*grade + 0.03*yr_built - 0.92*log_bathroom - 0.56*log_sqft_living + 0.03* log_sqft_lot + 0.59 log_sqft_above + 0.29 log_sqft_living15 - 0.33 log_sqft_lot15*

Then we use the test model and the predict() function to come up with a confusion table

```
newdataTrain.lda.values = predict(newdataTrain.lda,newdataTest)

tbl<-table(newdataTest$condition,newdataTrain.lda.values$class)

        1     2     3     4     5

  1     4     0     1     5     0
```

```
  2     4     2    25     6     0

  3     2     1  2500   203    65

  4     7     0   923   170    49

  5     4     0   217    79    56
```

To interpret the table, the diagonal values show the amount of correctly classified points. We developed a confusion table, using the command, and we get the percent of correct classification

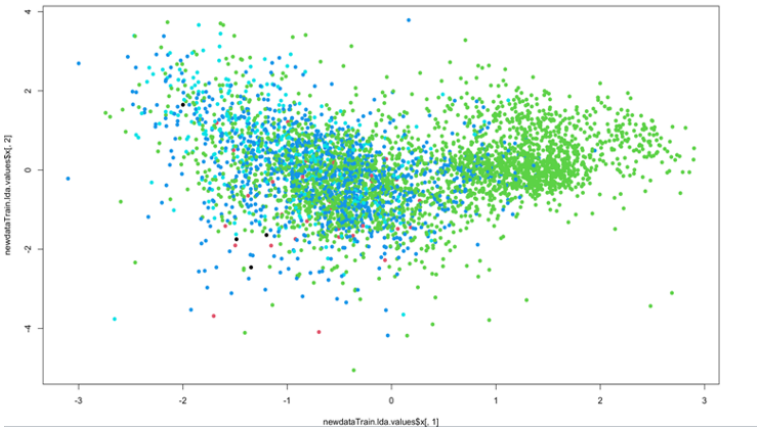
```
mean(newdataTrain.lda.values$class==newdataTest$condition)
[1] 0.6319685
```

Now we can look at some plots, we ran the following command:

```
> newdataTrain.lda.values$x
                LD1           LD2           LD3           LD4
2    -0.4482008056  0.8945971212 -1.118052e+00  4.562548e-01
3    -0.8831588775 -2.6284998293 -1.942898e+00 -2.567470e+00
8    -0.3933578800 -1.0670594666  9.238477e-02  1.155844e-01
11   -0.7657729611  1.0895351924  9.227208e-01 -5.282385e-01
21   -1.2863070536  0.8302170284  3.057360e-01  5.531195e-01
23    0.8955348754  0.1913102588  5.359241e-02  3.305867e-02
30    1.4196521641 -0.3812163271 -3.742040e-01 -7.292484e-01
35   -0.5149505007  0.6955016382  5.184323e-01 -4.788623e-01
```

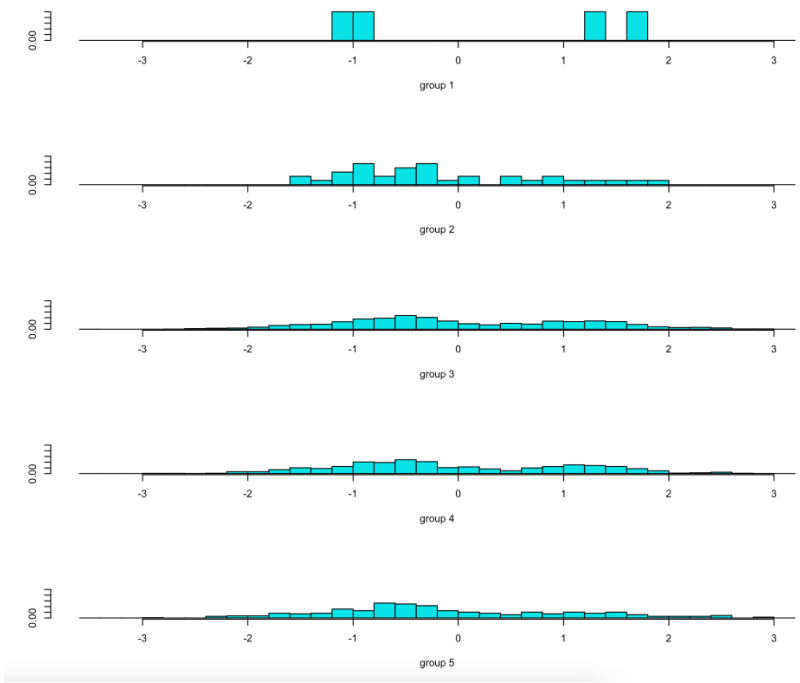
```
newdataTrain.lda.values$x
```

It gave us LD1, LD2, LD3, and LD4 data points. This is the plot it gave:



As you can see the data points overlap a lot and it is clearly hard to tell the separation of the colors blue, cyan, green, red, and black.

Now we will look at the Histogram plots for the LD1 data points:



To prove more, these are the groups very the numbers are the categories. You can see that they overlap a lot. As mentioned before, the goal here for LDA is to maximize the distance between 2 categories while minimizing the scatter. Here we still see bars overlapping others this tells us that this is the maximum amount of separation

Conclusion

The applications used to determine the predictors that have the maximum influence, based on the different types of regression used, we can determine that all the predictors are significant in determining the price of a house in King County (All the predictors after transforming several). We can also determine that OLS is a better model for the price as it produces the best accuracy leading to the waterfront having a stronger correlation and influence to price as its' parameter estimate was the largest at .4179 (if the house had a waterfront). As OLS produced a better model than regularized regression, the lambda values calculated in regularized regression were smaller than one indicating less improvement.

Using Principal factor analysis, we can group the variables into relevant categories, making the data easier to analyze and understand. The RC1 has all of the components that lead to a costly house, followed by the RC2 and RC3.

As mentioned before, the goal for LDA is to maximize the distance between 2 categories while minimizing the scatter. We still see bars overlapping others this tells that this is the maximum amount of separation. Although we were not able to get or interpret ROC.