MSBA 6120: Introduction to Statistics for Data Scientists

Predicting House Prices in Kings County: Project Whitepaper

17th August, 2019

UNIVERSITY OF MINNESOTA

Driven to Discover℠

Team 7: Section 2

Zheming Lian

Yassine Manane

Pahal Sushil Patangia

Rikarnob Bhattacharyya

Background of the project & Data Overview

The 2008 financial crisis was one of the worst in the history of mankind. A key driver of the crisis was something that analysts and financial experts called the “housing bubble”. Put in short, during the early 2000s there was a sharp decrease in interest rates for home loans. As Investopedia puts it, “The vast majority of loans were adjustable-rate mortgages with low initial rates”. The subsequent madness of buying homes was followed by house prices being driven up. When it became apparent to homeowners that prices might fall soon and they were living in a “bubble”, unprecedented selling of houses started, which drastically buried the prices, with people across the country increasingly finding themselves unable to pay off mortgages. And the rest is history.

House prices are a crucial moving part of every economy since it talks about the spending power of people and serves as a key economic indicator. For our project, we have chosen a dataset that provides us with housing prices and multiple other features/characteristics associated with a house that are potential drivers. Being able to identify the key drivers and subsequently predict house prices is what our project will aim to tackle.

Description of Datasets

We used a house pricing dataset sourced from Kaggle. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The dataset has 19 house features encompassing a good mix interval and nominal variables, plus the price and the id columns, along with 21,613 observations. An exhaustive list of fields is provided in the data dictionary in Appendix (table1). The prominent ones among them could potentially represent the following broad categories:

a. No. of bedrooms and bathrooms
b. No. of floors and area covered
c. House Rating
d. Zip code
e. Condition of the house
f. Amenities provided

Moreover, to better assess the impact of the house’s location on its price, we decided to combine the original dataset with another data source that links each zipcode with its city and the population density in that particular area. This approach will enable us to better interpret the results of our model.

Data Processing Summary

Before moving into the analysis, it is required that we perform certain data sanity checks to treat outliers and missing values. We assessed the univariate distributions and statistical plots of the dataset variables, we applied the following treatments for data cleaning. Further, we have created custom predictive characteristics that can incorporate the intermittent raw information more precisely for price prediction.

The details of the processing are enlisted below:

1. Cleanup of the timestamp values to convert the date field into a palatable format.
2. Mapping of zip codes to the corresponding populations to calculate the population density of the area.
3. The zip codes were also being mapped with their respective cities to gain a better understanding of price variations among different cities.
4. The recency of the house is also captured using it's renovation/origination information.
5. We found records where the number of bathrooms was not a whole number. This is because some of them are not full bathrooms. In general, a full bath contains four essential parts: a toilet, a sink, a bathtub, and a shower. A ¾ bath only contains three out of four parts, and a ½ bath contains only two out of four parts. To deal with this issue, we rounded such values to the nearest greatest integer under the assumption that certain integer reflects the number of all kinds of bathrooms in the house.

Descriptive Analysis

The distribution of price

The range of price is between $75,000 to $7,700,000. The mean price is $540,088, whereas the median is $450,000. The distribution of price is highly right-skewed, as 75% of the houses have a price under $645,000 and 95% of houses have a price under $1,156,480. (Figure 1)

The distribution of predictors

In general, the distribution of house features is highly skewed. For instance, while the maximum living area is 13,540 sq ft, the average living area is 2,080 sq ft, 75% of the houses have a living area of 2550 sqft, and 95% of the houses have a living area of 3,760 sq ft. (Figure 2). In terms of geographical information, 44% of the houses sold are in Seattle, Washington, while the second largest market of houses sold only takes up to 7% of the total house sell (Figure 4). This might be due to the imbalance of population size among cities as Seattle is the only metropolitan within King's County. In terms of demographic information, the distribution of population density is also skewed (Figure 5). In terms of the rating of the housing unit, both condition and grade fields have a relatively unskewed distribution (Figure 6).

(See Figure 3 for details of the distributions of other house features)

Visualization
We observed some plausible relationships between single house feature and sale price through visualization. In particular, visually:

1. Whether a house has a view of waterfront or not has an effect on the sale price (Figure 7).
2. Living area and sale price have a linear relationship (Figure 8).
3. The overall grade given to the house has an effect on the sale price (Figure 9).
4. Plot for PopDensity vs. price.
5. The location of the house in terms of the corresponding city has an effect on the sale price (Figure 10).

Outliers and Missing data
We found and removed one potential outlier with 33 bedrooms and 1620 sq ft of the living area (Figure 11). Our data did not have any missing information (Figure 12).

Inferential Analysis

Scope of inference
The unit of analysis is a house sold between May 2014 and May 2015, in King County, Washington state.

Model selection Criteria
The final model was arrived in an iterative fashion after considering the impact of various predictors starting from the 'kitchen-sink' process.

The selection of the final model is based on certain criteria:

1. Logical predictors: We only include a predictor when it has a logical relationship with Price.
2. Significant predictors: We take out the predictor with the highest p-value until reach a threshold (0.05 in our case).
3. Interpretability: We expect the model to be palatable. Following this principle, we removed the"zipcode" predictor because it has too many levels making it difficult to interpret.
4. Simplicity: Take out predictors when the model performance in terms of adjusted R^2 is not reduced vastly.
5. Maximized adjusted R^2: we always look for a model with the highest adjusted R^2 as long as other criteria are not violated.

Modeling
Based on criteria as listed above, our final model contains the following predictors:

1. City: nominal variable, indicates the location of the house
2. Waterfront: nominal variable, indicates whether a house has a view of waterfront or not
3. Grade: ordinal variable, indicates the overall grade of the house, based on King County grading system

4. Sqft_living: interval variable, indicates the square footage of home
5. PopDens: interval variable, indicates the population density of the zip code that corresponds to the house

Note: At the time of presentation we also included the "bathroom" predictor in the model. However, we found that the correlation between 'bathroom' and 'Sqft_living' is over 0.8. Given that the removal of 'bathroom' does not affect the adjusted R^2, we have decided to drop that variable.

The final model has the capacity to explain 75% of the variation in house prices. (Figure 13)

Model explanation
In terms of interval predictors, we observed from the correlation matrix of interval predictors that Sqft_living and PopDens are moderately correlated (Figure 14). Therefore, we are able to interpret that keeping other factors fixed, one additional unit of sqft_living on average will lead to an increase in sale price by $168.2. Moreover, keeping other factors fixed, one unit of increase in population density will lead to an increase in sale price by $3.4.

In terms of ordinal and nominal predictors, some levels of certain predictors have a significantly different impact on price, compared with the baseline level of those predictors, under the alpha level of 0.001. Therefore, we concluded that there are relationships between grade and price, between the corresponding city and price, and between the presence of waterfront and price.

Model diagnosis and assumption check
As we did not observe any explicit pattern on the residual plot of our final model (Figure 15), we concluded that we don't have a non-linear relationship that is not captured by the model.

Recommendation

To conclude, it turns out that a house's price is particularly sensitive to its area, grade and location. In particular, people interested in purchasing a house in Seattle may find it more interesting to invest in a house in the nearby cities for the same area and grade but with a considerably lower price. Similarly, one can opt for a house in a less dense area, not only it is less stressful, but it also enables some cost savings. For more upscale living, Mercer Island and Medina would be good choices - low population density and high-grade houses provides for luxurious options.

Lastly, even though this analysis concerns essentially the King's County, the same reasoning may be scaled to predict house prices in other regions.

Appendix

Table 1. Data Dictionary

Field Name	Description
id	notation for a house
date	Date house was sold
price	Price is the prediction target
bedrooms	Number of Bedrooms/House
bathrooms	Number of bathrooms/House
sqft_living	square footage of the home
sqft_lot	square footage of the lot
floors	Total floors (levels) in house
waterfront	House which has a view to a waterfront
view	Has been viewed
condition	How good the condition is (Overall)
grade	The overall grade given to the housing unit, based on King County grading system
sqft_above	square footage of house apart from the basement
sqft_basement	square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	zip
lat	Latitude coordinate
long	Longitude coordinate
sqft_living15	Living room area in 2015 (implies-- some renovations) This might or might not have affected the lot size area

sqft_lot15	lotSize area in 2015 (implies-- some renovations)

Figure 1. Price distribution

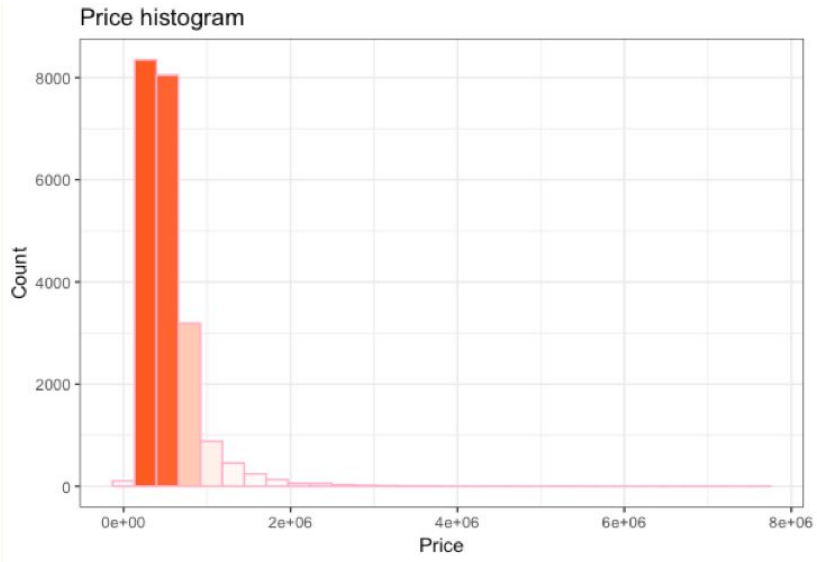


Figure 2. Distribution of sqft_living

```
summary(house_new2$sqft_living)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  290    1400    1880    2045    2500   13540
```

Figure 3. Distribution of other house features

```
summary(house_new2$sqft_lot)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  520    5000    7500   14706   10350 1651359
summary(house_new2$sqft_above)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  290    1180    1530    1747    2140    9410
summary(house_new2$sqft_basement)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.0     0.0     0.0   297.9   580.0  4820.0
summary(house_new2$bedrooms)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   3.000   3.000   3.351   4.000  11.000
```

```
summary(house_new2$bathrooms)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   2.000   3.000   2.411   3.000   8.000
```

Figure 4. Proportion of corresponding cities

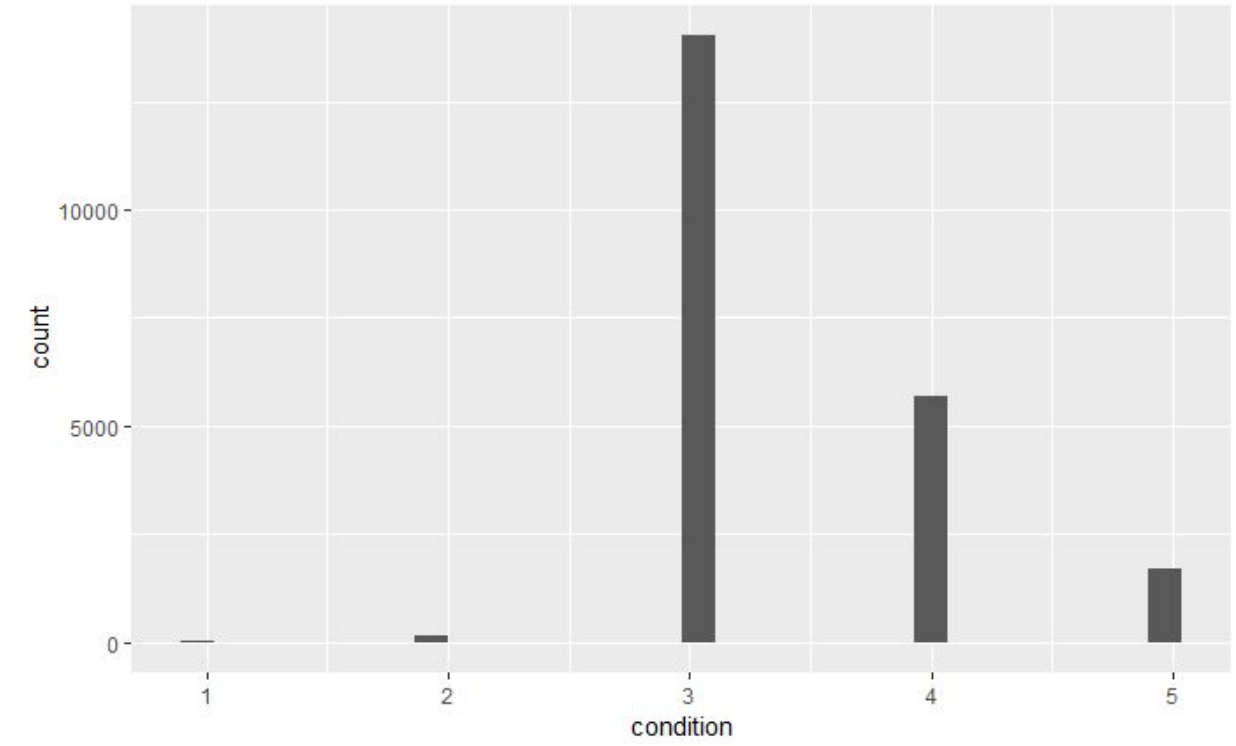
```
> sort(table(house_new2$City),decreasing = TRUE)/dim(house_new2)[1]

      Seattle, Washington         Renton, Washington
                0.440907751                0.078445820
     Bellevue, Washington        Redmond, Washington
                0.069112879                0.048089203
     Kirkland, Washington           Kent, Washington
                0.047990962                0.046517340
       Auburn, Washington   Federal Way, Washington
                0.044798114                0.038265056
     Issaquah, Washington  Maple Valley, Washington
                0.036005502                0.028981236
   Snoqualmie, Washington        Kenmore, Washington
                0.015227429                0.013901169
Mercer Island, Washington   Woodinville, Washington
                0.013852048                0.013409962
     Enumclaw, Washington    North Bend, Washington
                0.011494253                0.010855683
      Bothell, Washington         Duvall, Washington
                0.009578544                0.009332940
    Carnation, Washington         Vashon, Washington
                0.006090972                0.005796247
Black Diamond, Washington     Fall City, Washington
                0.004912074                0.003978780
       Medina, Washington
                0.002456037
```

Figure 5. Distribution of population density

```
summary(house_new2$PopDens)
  Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 66.56  1408.91  3756.89  4016.47  6041.66 14594.02
```

Figure 6. Distribution of third party rating (Condition, Grade)



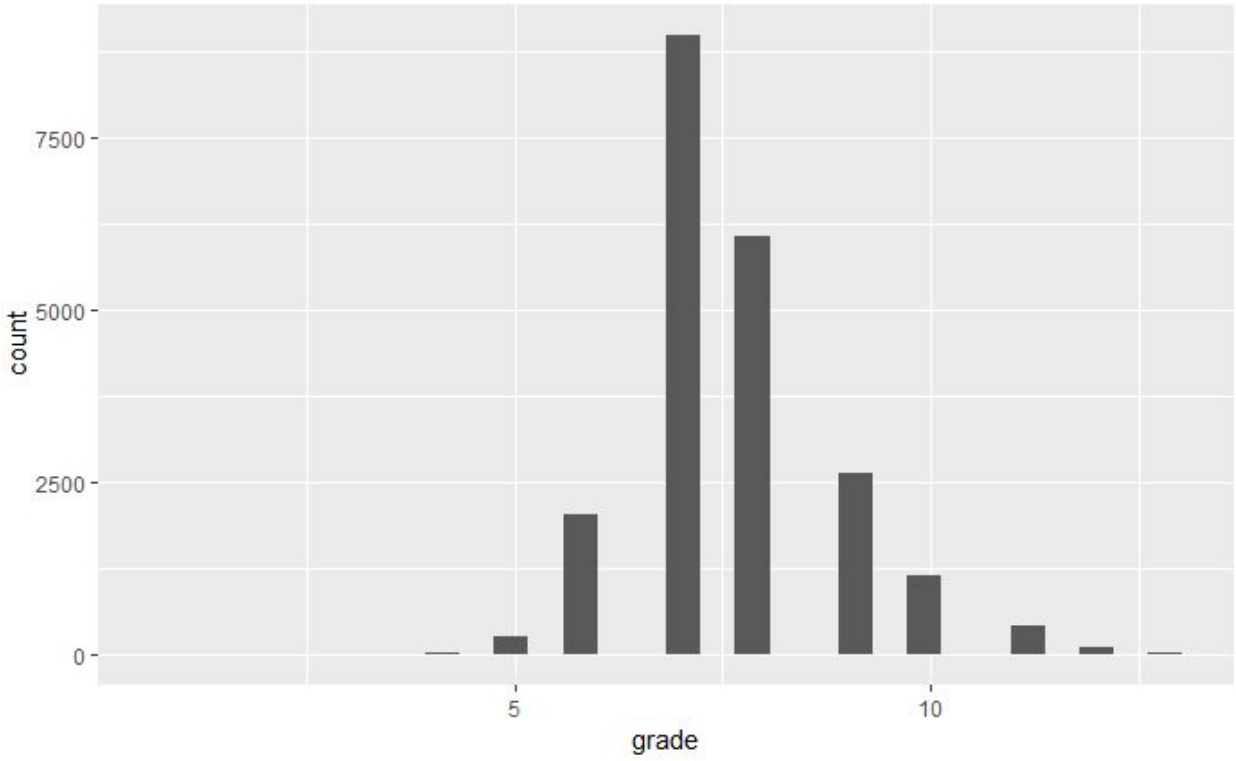


Figure 7. Waterfront vs. Price

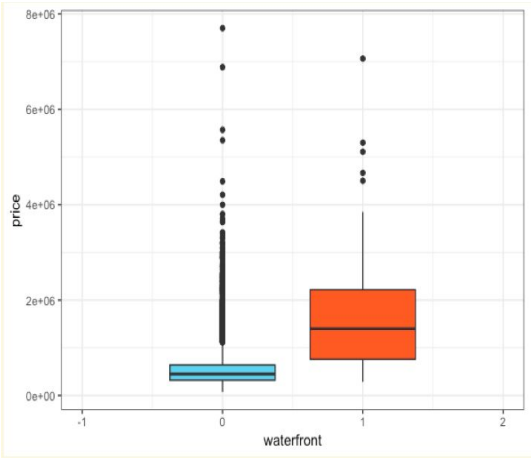


Figure 8. Living room area vs. Price

Figure 9. Grade vs. Price

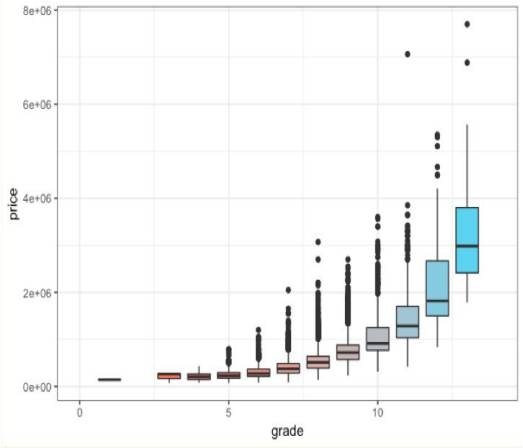


Figure 10. Distribution of price across different cities

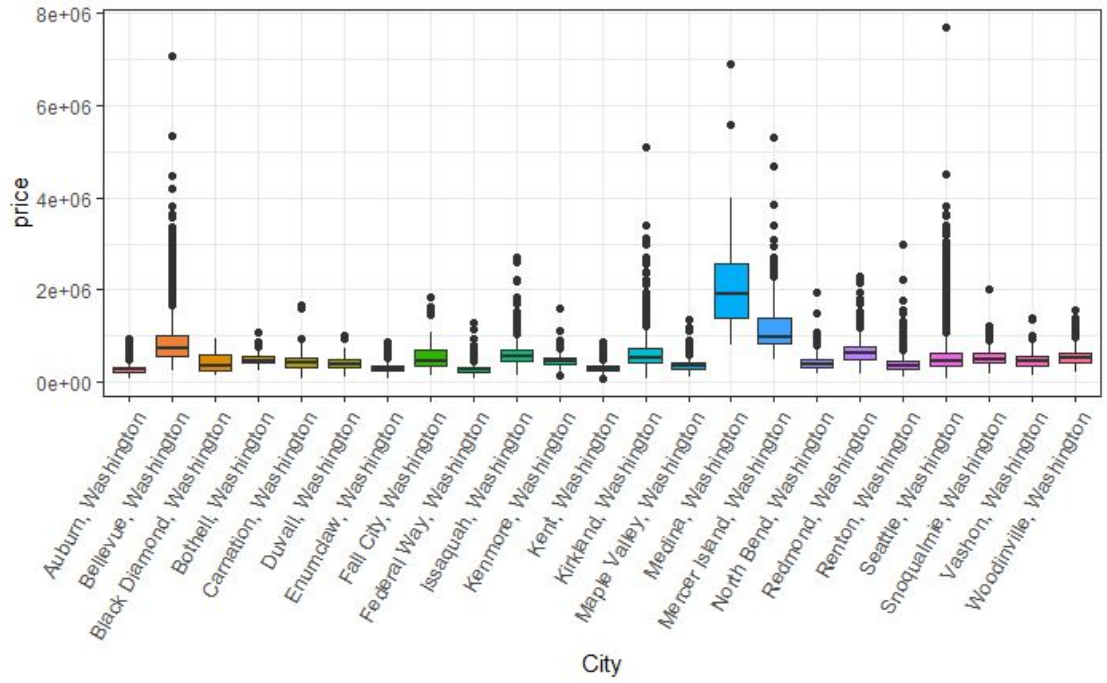


Figure 11. Outlier record

```
> house_new2[house_new2$bedrooms == 33,]
             id     date  price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above
15871 2402100895 20140625 640000       33      1.75        1620     6000      1          0    0         5     7       1040
      sqft_basement yr_built yr_renovated zipcode sqft_living15 sqft_lot15 total_sqft total_sqft15 change_in_sqft year_sold
15871           580     1947            0   98103          1330       4700       8200         6610          -1590      2014
      newness bed_bath_ratio   date_new area_floor_ratio area_floor_ratio_15 area_floor_change q75
15871      67       18.85714 2014-06-25             1620                1330              -290   0
```

Figure 12. # of NA for each column (No NA in the data)

```
```{r}
colSums(is.na(house))
```
```

```
          id          date         price      bedrooms     bathrooms   sqft_living      sqft_lot        floors    waterfront
           0             0             0             0             0             0             0             0             0
        view     condition         grade    sqft_above sqft_basement      yr_built  yr_renovated       zipcode           lat
           0             0             0             0             0             0             0             0             0
        long sqft_living15    sqft_lot15
           0             0             0
```

Figure 13. Summary table of the final regression model

```
Call:
lm(formula = price ~ as.factor(City) + as.factor(waterfront) +
    sqft_living + as.factor(grade) + PopDens, data = house_new2)

Residuals:
     Min       1Q   Median       3Q      Max
-1557877   -89869    -6895    69184  3794954

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.427e+05  1.861e+05  -0.767  0.44302
as.factor(City)Bellevue, Washington      3.114e+05  8.094e+03  38.480  < 2e-16 ***
as.factor(City)Black Diamond, Washington 1.257e+05  1.952e+04   6.437 1.24e-10 ***
as.factor(City)Bothell, Washington       7.442e+04  1.465e+04   5.081 3.80e-07 ***
as.factor(City)Carnation, Washington     1.601e+05  1.777e+04   9.008  < 2e-16 ***
as.factor(City)Duvall, Washington        1.412e+05  1.479e+04   9.545  < 2e-16 ***
as.factor(City)Enumclaw, Washington      1.078e+05  1.363e+04   7.911 2.69e-15 ***
as.factor(City)Fall City, Washington     2.313e+05  2.161e+04  10.703  < 2e-16 ***
as.factor(City)Federal Way, Washington  -1.046e+05  9.263e+03 -11.297  < 2e-16 ***
as.factor(City)Issaquah, Washington      2.015e+05  9.291e+03  21.692  < 2e-16 ***
as.factor(City)Kenmore, Washington       7.755e+04  1.265e+04   6.129 9.00e-10 ***
as.factor(City)Kent, Washington          2.239e+03  8.586e+03   0.261  0.79424
as.factor(City)Kirkland, Washington      1.969e+05  8.725e+03  22.566  < 2e-16 ***
as.factor(City)Maple Valley, Washington  6.727e+04  9.810e+03   6.858 7.20e-12 ***
as.factor(City)Medina, Washington        1.191e+06  2.712e+04  43.916  < 2e-16 ***
as.factor(City)Mercer Island, Washington 4.459e+05  1.287e+04  34.641  < 2e-16 ***
as.factor(City)North Bend, Washington    1.613e+05  1.394e+04  11.571  < 2e-16 ***
as.factor(City)Redmond, Washington       1.997e+05  8.576e+03  23.285  < 2e-16 ***
as.factor(City)Renton, Washington        5.032e+04  7.690e+03   6.543 6.15e-11 ***
as.factor(City)Seattle, Washington       1.054e+05  7.473e+03  14.103  < 2e-16 ***
as.factor(City)Snoqualmie, Washington    1.410e+05  1.225e+04  11.505  < 2e-16 ***
as.factor(City)Vashon, Washington        1.114e+05  1.841e+04   6.052 1.46e-09 ***
as.factor(City)Woodinville, Washington   1.779e+05  1.280e+04  13.896  < 2e-16 ***
as.factor(waterfront)1                   7.116e+05  1.574e+04  45.198  < 2e-16 ***
sqft_living                              1.682e+02  2.284e+00  73.650  < 2e-16 ***
as.factor(grade)3                        1.174e+05  2.146e+05   0.547  0.58437
as.factor(grade)4                        5.306e+04  1.892e+05   0.281  0.77909
as.factor(grade)5                        9.727e+03  1.863e+05   0.052  0.95837
as.factor(grade)6                       -5.351e+03  1.860e+05  -0.029  0.97705
as.factor(grade)7                        1.435e+04  1.860e+05   0.077  0.93850
as.factor(grade)8                        5.847e+04  1.860e+05   0.314  0.75332
as.factor(grade)9                        1.644e+05  1.861e+05   0.884  0.37696
as.factor(grade)10                       3.476e+05  1.862e+05   1.867  0.06196 .
as.factor(grade)11                       5.910e+05  1.865e+05   3.169  0.00153 **
as.factor(grade)12                       9.791e+05  1.875e+05   5.222 1.78e-07 ***
as.factor(grade)13                       2.109e+06  1.937e+05  10.887  < 2e-16 ***
PopDens                                  3.423e+01  8.227e-01  41.611  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 184800 on 20322 degrees of freedom
Multiple R-squared:  0.7511,    Adjusted R-squared:  0.7507
F-statistic:  1704 on 36 and 20322 DF,  p-value: < 2.2e-16
```

Figure 14. Correlation matrix of interval predictors in the final model

	sqft_living	PopDens
sqft_living	1.0000000	-0.1816125
PopDens	-0.1816125	1.0000000

Figure 15. Residual plot of the final model