Statistical Analysis & Modeling of King County Housing Market

Vaishnavi Udaya Kumar

Net Id: MA5458

Supervisor: Prof. Jiyoun Myung

A Final Project Report for STAT 632 (Linear and Logistic Regression) submitted to the Dept. of Statistics & Biostatistics, California State University, East Bay in partial fulfilment of the requirements for the degree of

Masters in Statistics (Data Science Concentration)

Spring 2021

Table of Contents

Abstract

Statistical Analysis & Modeling of King County Housing Market
Vaishnavi Udaya Kumar, California State University, East Bay

The real estate market plays an important role not just in determining a region's development, but also has a huge impact on the economy. Therefore, it becomes all the more important to gather and study real estate data closely to better understand the various factors which can affect property prices and sales.

My findings show that we can establish a linear relationship between the home price and predictors such as number of bedrooms, bathrooms, property square footage, condition etc. Using this premise, I have been able to develop an MLR model for predicting property prices and a Logistic Regression model to predict the bidding direction. Lastly, I have attempted to choose profitable zip codes for real estate agents to maximize their revenue.

Problem and Motivation

The real estate market is an important indicator of a given city's, county's or state's well-being, it's citizens' quality of living, and as we had seen little over a decade ago in 2008, it can have profound implications on the larger economic market as well. Due to the volatile nature of the housing market, the ever changing topography of cities, and the numerous factors which can impact home prices, it is in our interest to study such data, identify trends, meaningful correlations and key drivers of property sales.

The motivation to choose this data set, was to be able to apply the statistical concepts and techniques of Linear and Logistic Regression learnt in the STAT 632 course, to be able to analyze and model the data, and derive useful insights. Likewise, today we see companies such as Zillow, Realogy, Redfin etc., being heavily dependent on Data Science & Machine Learning to gain an edge over their competitors in important markets. This project is also an attempt to try and understand how one can leverage the insights from our experiments to support the real-estate business from a financial perspective in a data-driven market.

Data Description

0.1 Data Source

The primary data set selected for the project pertains to Home Sales in King County, USA. It is the most populous county in Washington, and the 12th most populous in the United States. This data set includes homes sold between May 2014 and May 2015, and is sourced from Kaggle.

SNOHOMISH COUNTY
KING COUNTY
SHORELINE
LAKE FOREST PARK
KENMORE
405
KIRKLAND
Sammamish River
REDMOND
Snoqualmie River
Puget Sound
5
Lake Washington
520
SEATTLE
Elliott Bay
BELLEVUE
SAMMAMISH
Lake Sammamish
PORT OF SEATTLE
MERCER ISLAND
Duwamish Waterway
ISSAQUAH
90
SNOQUALMIE
BURIEN
TUKWILA
RENTON
NORMANDY PARK
PORT OF SEATTLE
KING COUNTY
18
Cedar River
KENT
5

0.2 Data Features

This data set consists of 21,613 observations on the following 21 variables:

- **id** - Unique ID used as a notation for each property sold
- **date** - Date of the property sale
- **price** - Price of each property sold
- **bedrooms** - Number of bedrooms in the property
- **bathrooms** - Number of bathrooms (where .5 accounts for a room with a toilet but no shower)
- **sqft_living** - Square footage of the home interior living space
- **sqft_lot** - Square footage of the land space

- **floors** - Number of floors or storeys
- **waterfront** - A dummy variable for whether the property was overlooking the waterfront or not
- **view** - A 0-4 rating on the view from the property.
- **condition** - A 1-5 rating on the condition of the property.
- **grade** - An index from 1 to 13, where 1-3 fall short of building construction and design, 4-7 have an average level of construction and design, 8-11 have a good quality 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** - Latitude coordinates of the property
- **long** - Longitude coordinates of the property
- **sqft_living15** - The avg. square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15** - The avg. square footage of the land lots of the nearest 15 neighbors
- Additionally, we add variables namely:
  - **is_renovated**: serves as a flag to identify whether a house was renovated.
  - **is_basement**: serves as a flag to identify whether a house has a basement.
  - **avg_price_per_zip**: Avg. price of homes in a given zip code.
  - **above_avg_zip_price**: Flag to indicate if a home's value is above (1) or below (0) the avg home price in that zip code.

Questions of Interest

1.1 Research Question 1

”Analyzing the data with statistical techniques, will we be able to explain the relationship between the various features of housing data and the price?”

Here, I am attempting to investigate if there is a linear relationship between the various features of housing data, and by inference, predict the price of a given house from it’s key features.

1.2 Research Question 2

“As a real estate agent, can I help my clients to place their bids accurately above or below the asking price while buying homes in specific locations using properties’ attributes with, proven reliability?”

With this question, I am attempting to develop a Logistic Regression model which can be trained on historical data to accurately predict the bidding direction i.e., whether to bid over or under the asking price of a given house, using it’s attributes. To satisfy the last part of the questions i.e., reliability, I plan to set aside a part of the data for testing and measure the model’s efficacy.

1.3 Research Question 3

“Can we help real estate agents maximize their business by suggesting profitable zip codes to buy/sell homes in?”

While the first two questions focused on the modeling aspect, I attempt to answer the last question from a data analysis standpoint. As we know, the national average for agent commission is largely constant, and hence their revenue is dependent on the homes’ value. By analyzing the housing data, if I can correctly identify significantly more expensive neighborhoods, we can equip agents with this information to help focus their business on these zip codes and thereby maximize their revenue.

3. Regression Analysis, Results and Interpretation

2.1 Exploratory Data Analysis

2.1.1 Summary

- **id** - 21,436 unique ids are available. About 177 duplicate ids were found, which could be the small percentage of homes sold within the duration of one year. We can removed the duplicate ids by only retaining the properties with the most recent date since only these would be relevant to our study.

- **date** -

  Min: 2014-05-01
  Maximum: 2015-05-01

- **price** -

  Minimum: $75,000
  Median: $450,000
  Maximum: $7,700,000

- **bedrooms** -

  Minimum: 0 (Studio)
  Median: 3 bedrooms
  Maximum: 33

- **bathrooms** -

  Minimum: 0
  Median: 2.25
  Maximum: 8

- **sqft_living** -

  Minimum: 290
  Median: 1910
  Maximum: 13540

- **sqft_lot** -

  Minimum: 520
  Median: 7,617
  Maximum: 1,651,359

- **floors** -

  Minimum: 1
  Median: 1.5
  Maximum: 3.5

- **condition** -

  Minimum: 1
  Median: 3
  Maximum: 5

- **grade** -

  Minimum: 1
  Median: 7
  Maximum: 13

- **sqft_above** -

  Minimum: 290
  Median: 1560
  Maximum: 9410

- **yr_built** -

  Minimum: 1900
  Median: 1975
  Maximum: 2015

- **sqft_living15** -

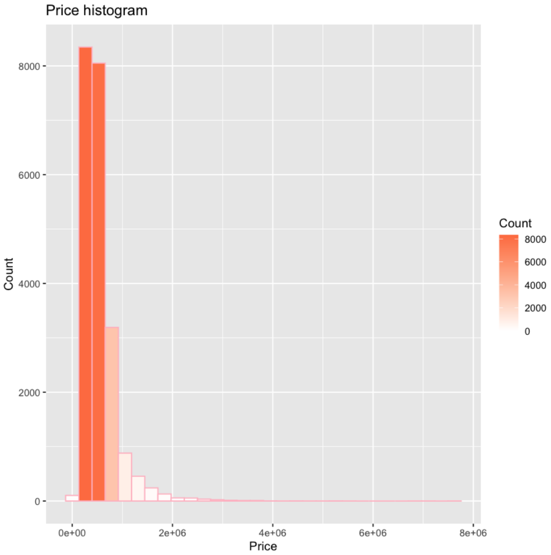
  Minimum: 399
  Median: 1840
  Maximum: 6210

- **sqft_lot15** -

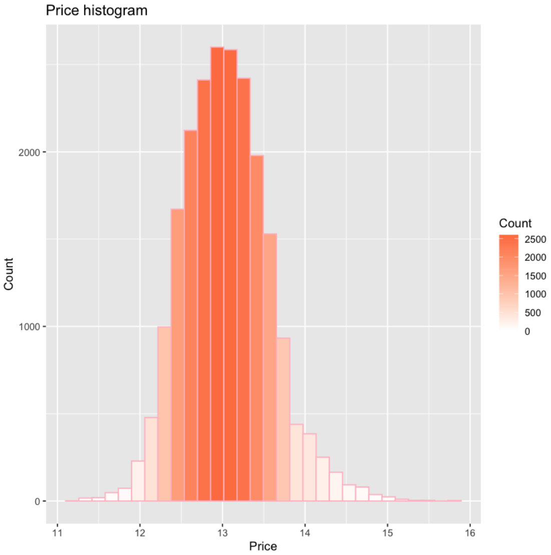
  Minimum: 651
  Median: 7620
  Maximum: 871,200

The first two features **is_renovated** & **is_basement** are used for both Research Questions 1 & 2. The last two features **avg_price_per_zip** & **above_avg_zip_price** are only used for Research Question 2.

Let us take a closer look at **price**, our response variable:



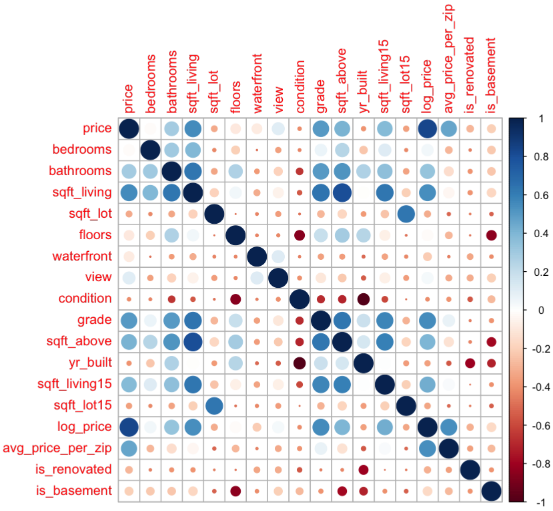
We see a significant left skew with our response. Let us try to normalize this by considering log transformation on price.



As we can see, log tranformation of price provides a normal distribution of our response. For further EDA, let us consider **log(price)** as our response.

2.1.2 Correlation

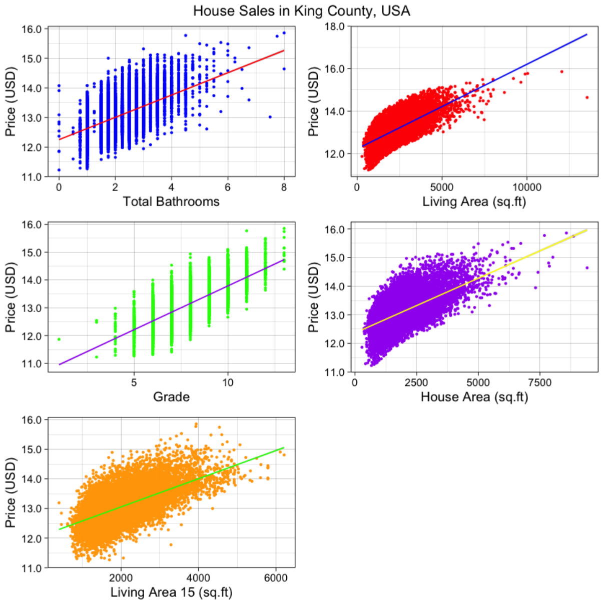
Let us look at how our predictors are correlated with our response i.e., **price**



From the correlation plot, we see that:

- bathrooms
- sqft_living
- grade
- sqft_above
- sqft_living15
- avg_price_per_zip

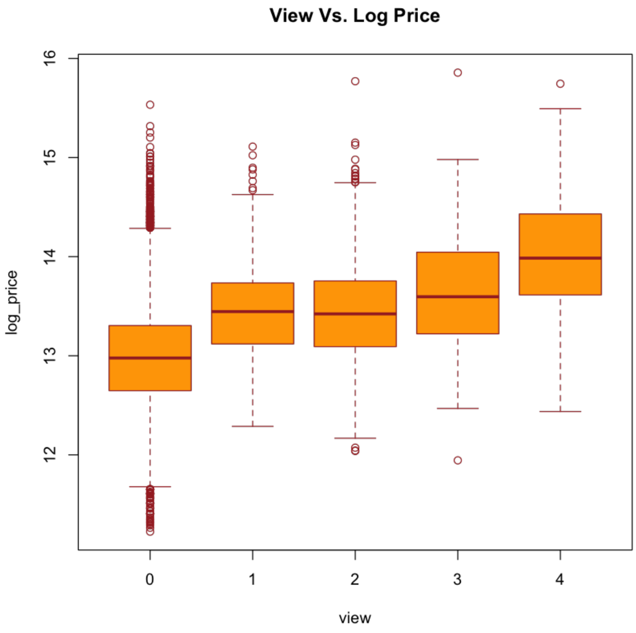
have a positive correlation with the response. Next, let us now see how these predictors are distributed with respect to response.



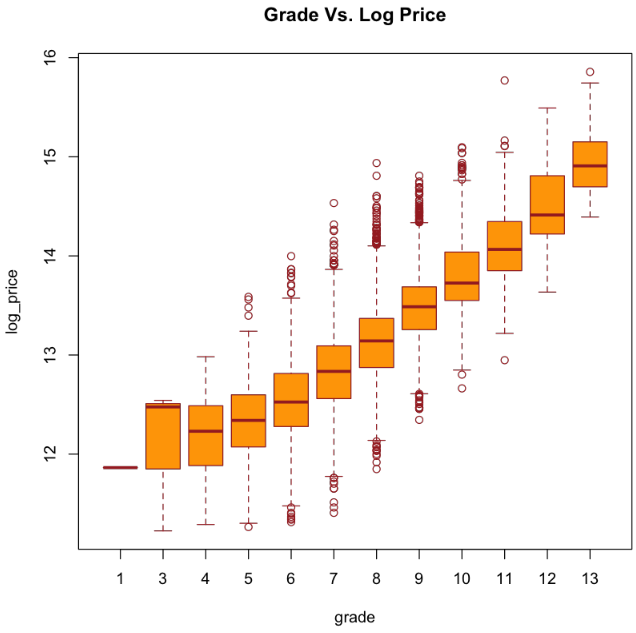
We see a pronounced linear trend between the aforementioned variables and response.

Now, let us look at the discrete and categorical features, and their relationship with the response:

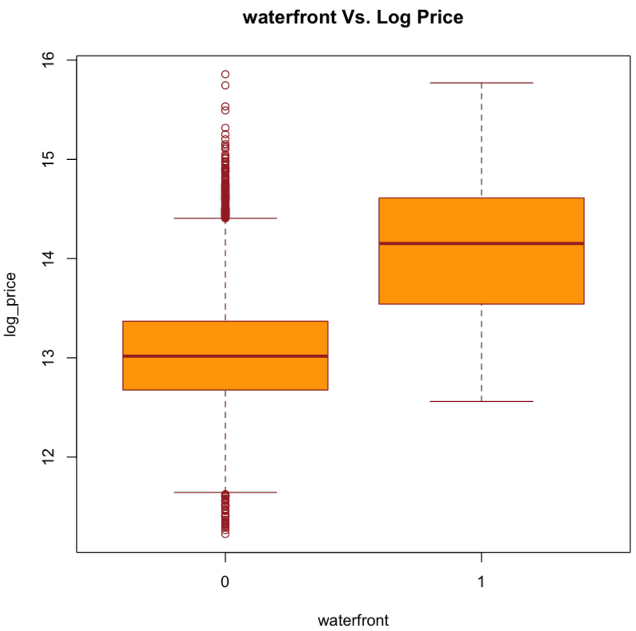
- **View**:



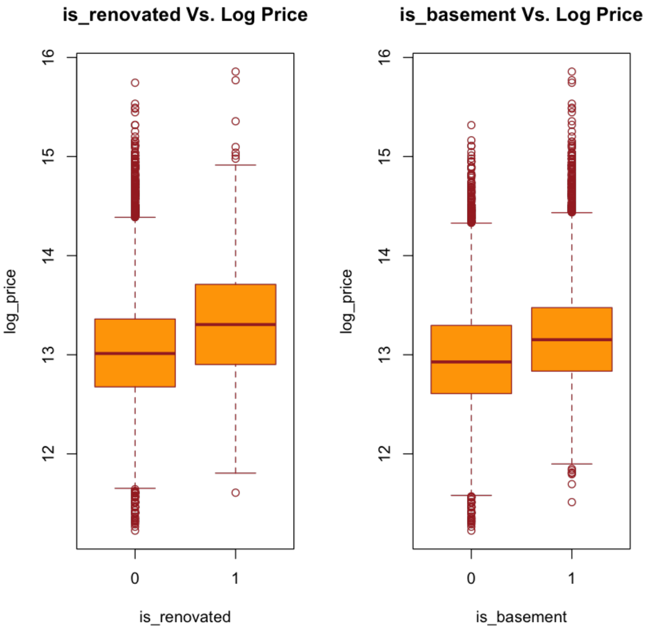
- **Grade**:



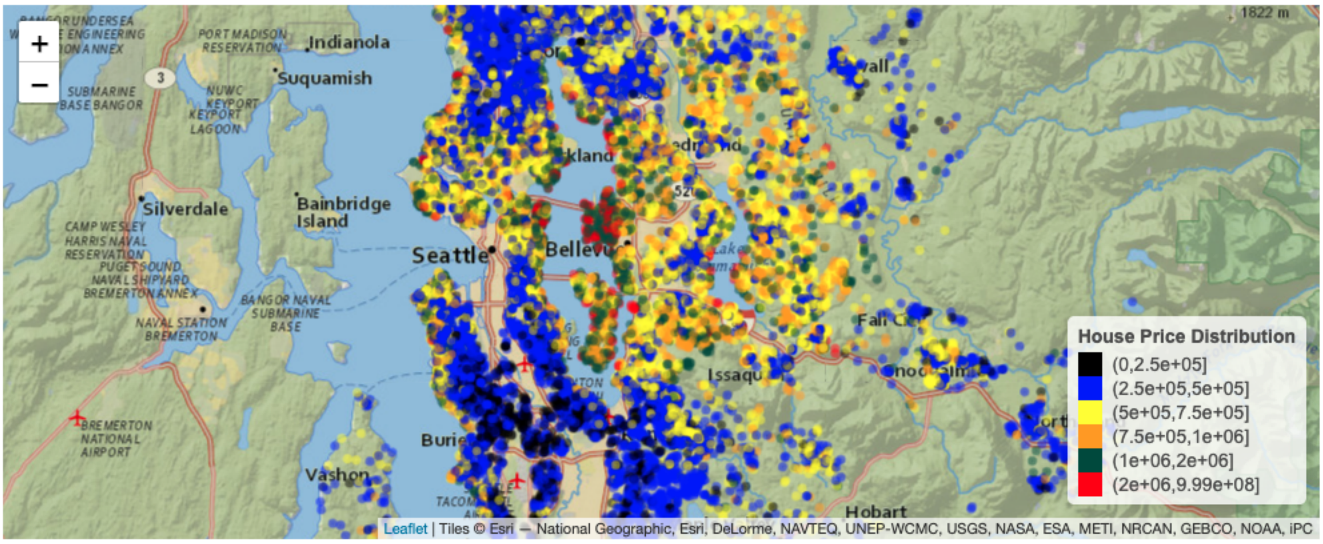
- **Waterfront**:



- **is_renovated & is_basement**:



- Distribution of homes in King County by price



2.2 Important Details of the Analysis

From our exploratory data analysis, we see that our dataset contains both linear and non-linear data. While we are able to use the above features in the previous section for our modeling, it is important that we identify the non-linear features and exclude it from our modeling exercise. Looking at the features in our dataset, at the outset, we can identify features such as:

- id
- date
- zip code
- lat
- long

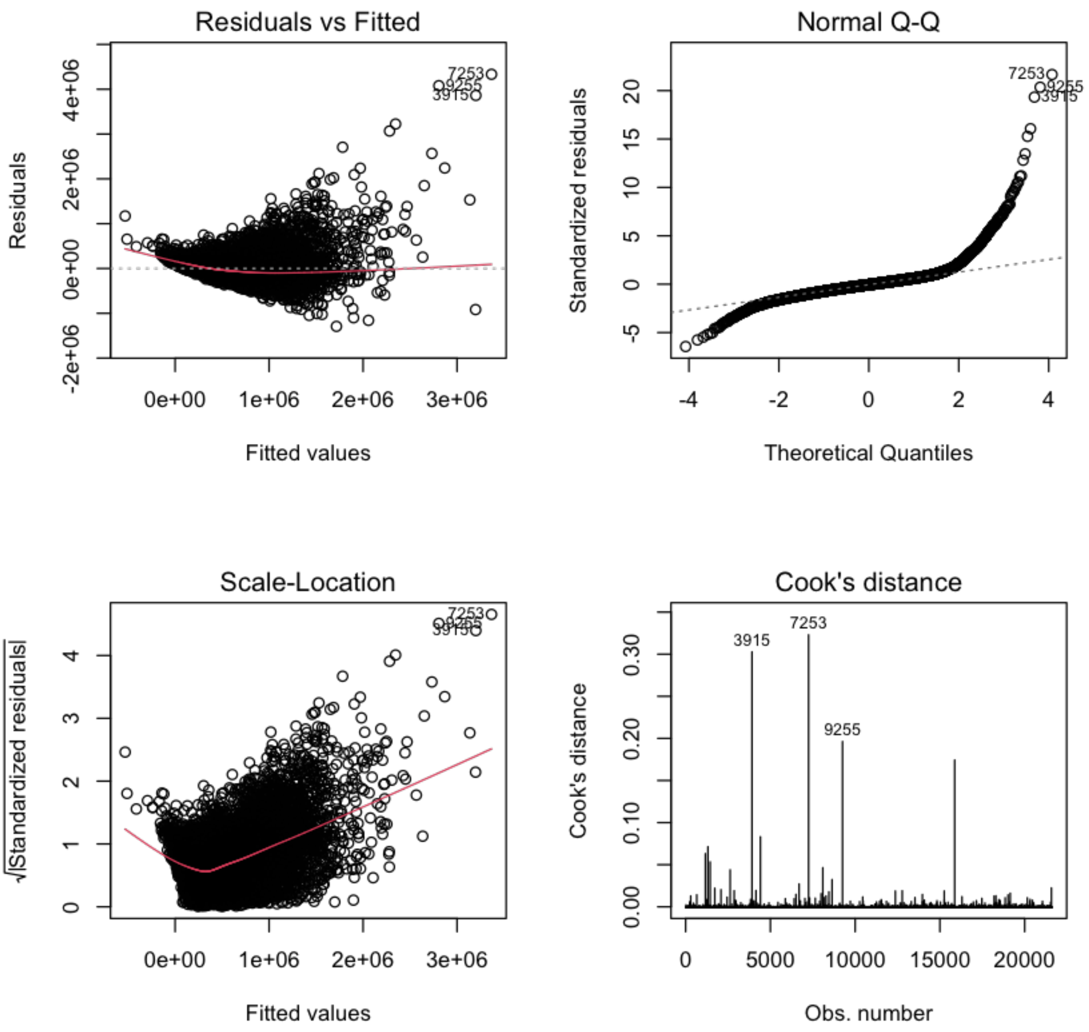
which are clearly non-linear. We can also confirm this by testing whether the assumptions of linearity are satisfied.

Further analysis discussed in further sections along with respective Research Question.

2.3 Diagnostic Checks

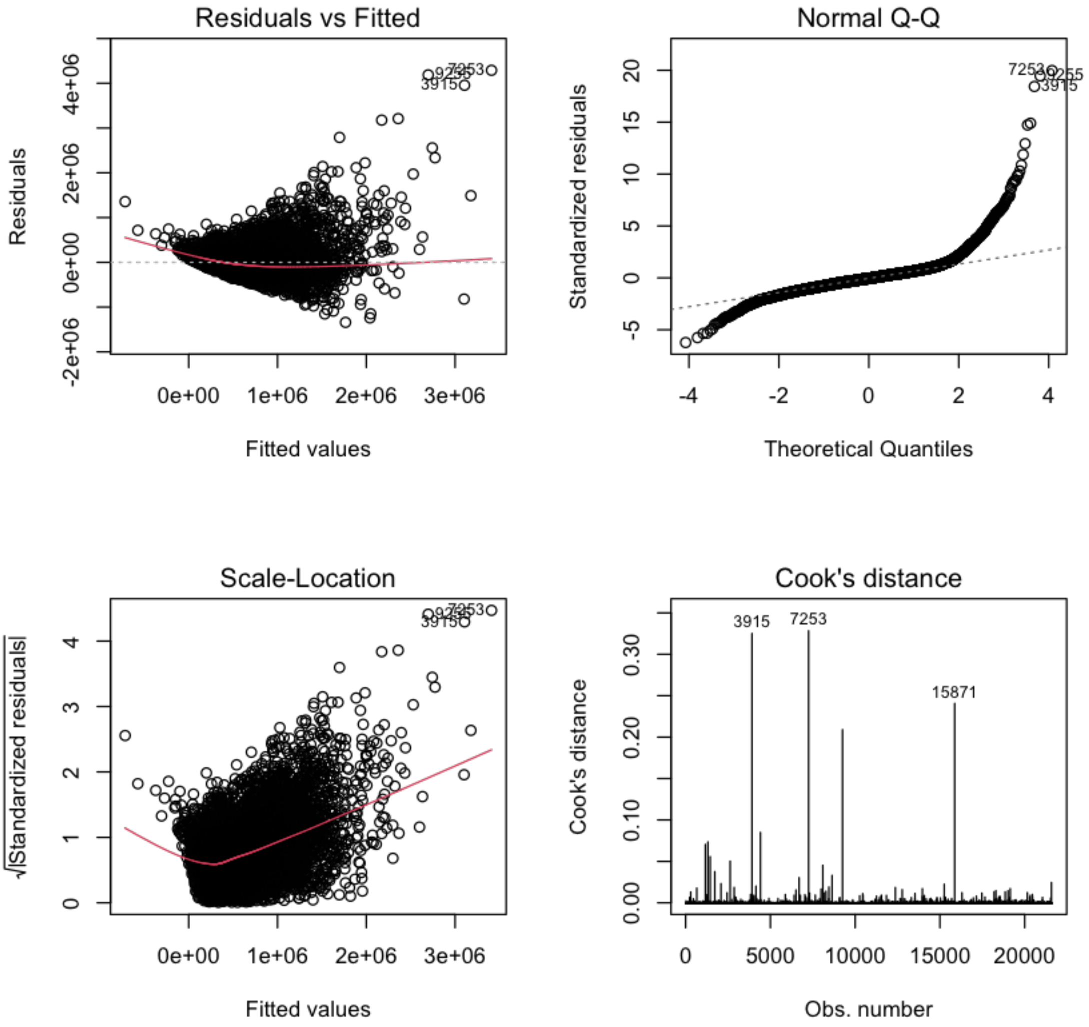
2.3.1 For Research Question 1

- Full model
  - First, let us test the data, as-is for linearity. Here, we use **price** as the response and all the features except **id** & **date** as predictors.



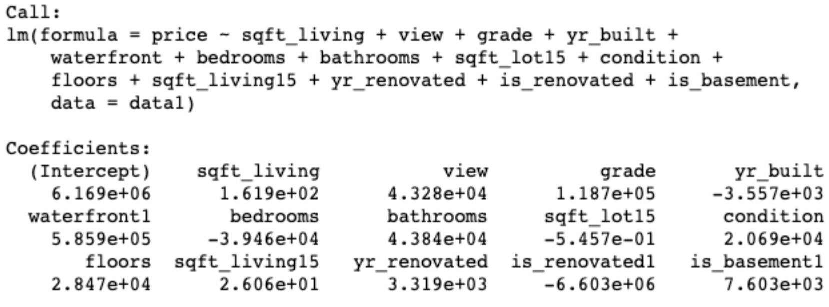
Here, we see clearly that the assumptions of MLR are not satisfied. Especially, the Normal Q-Q plot does not appear to be linear. Hence, we cannot use these features, as-is for linear modeling.

  - Next, let us test the data, by removing the features which seem to be inherently non-linear (as we saw in **Section 2.2**). Here, we use **price** as the response and all the features except **id** & **date**, **zipcode**, **lat**, **long** as predictors.

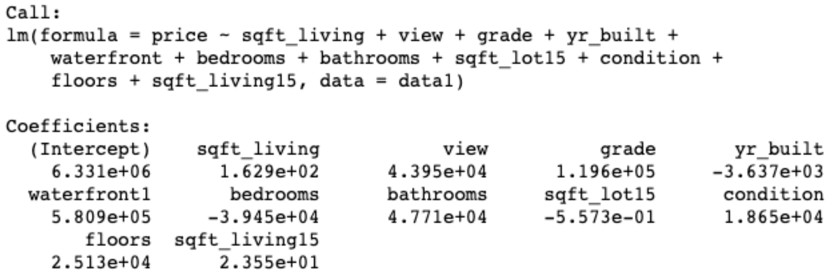


Here too, we see clearly that the assumptions of MLR are not satisfied. and the Normal Q-Q plot still does not appear to be linear.

- Reduced model using Variable Selection
- Now, let us use Variable Selection to find the right subset of predictors so as to satisfy the assumptions of MLR and fit our model. We employ the following methods:
  - Forward AIC & BIC

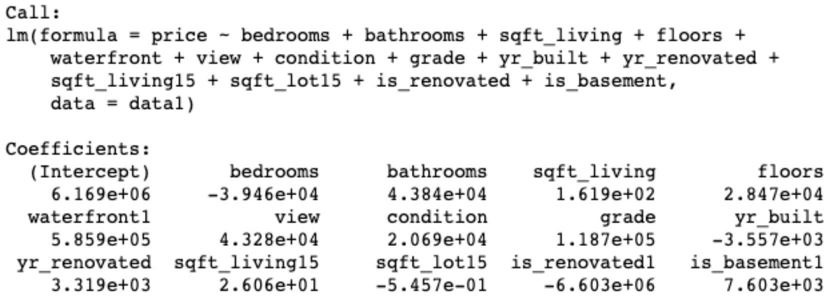


```
Call:
lm(formula = price ~ sqft_living + view + grade + yr_built +
    waterfront + bedrooms + bathrooms + sqft_lot15 + condition +
    floors + sqft_living15 + yr_renovated + is_renovated + is_basement,
    data = data1)

Coefficients:
  (Intercept)    sqft_living           view          grade       yr_built
    6.169e+06      1.619e+02      4.328e+04      1.187e+05     -3.557e+03
  waterfront1       bedrooms      bathrooms     sqft_lot15      condition
    5.859e+05     -3.946e+04      4.384e+04     -5.457e-01      2.069e+04
       floors  sqft_living15   yr_renovated  is_renovated1   is_basement1
    2.847e+04      2.606e+01      3.319e+03     -6.603e+06      7.603e+03
```

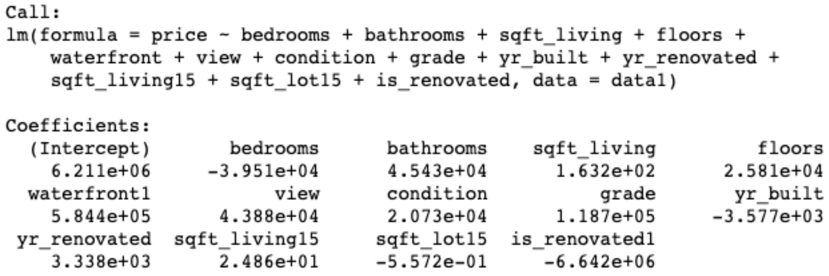


```
Call:
lm(formula = price ~ sqft_living + view + grade + yr_built +
    waterfront + bedrooms + bathrooms + sqft_lot15 + condition +
    floors + sqft_living15, data = data1)

Coefficients:
  (Intercept)    sqft_living           view          grade       yr_built
    6.331e+06      1.629e+02      4.395e+04      1.196e+05     -3.637e+03
  waterfront1       bedrooms      bathrooms     sqft_lot15      condition
    5.809e+05     -3.945e+04      4.771e+04     -5.573e-01      1.865e+04
       floors  sqft_living15
    2.513e+04      2.355e+01
```

∗ Backward AIC & BIC

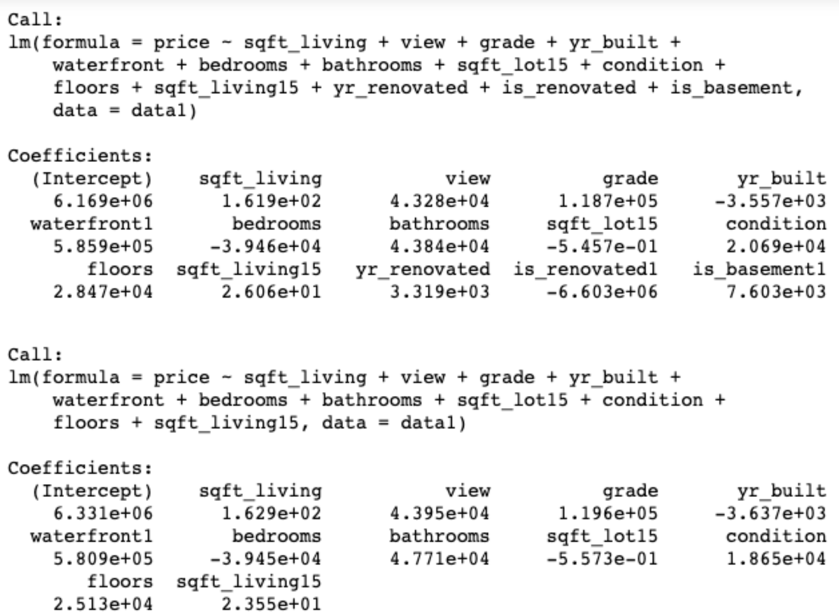


```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + floors +
    waterfront + view + condition + grade + yr_built + yr_renovated +
    sqft_living15 + sqft_lot15 + is_renovated + is_basement,
    data = data1)

Coefficients:
  (Intercept)       bedrooms      bathrooms    sqft_living         floors
    6.169e+06     -3.946e+04      4.384e+04      1.619e+02      2.847e+04
  waterfront1           view      condition          grade       yr_built
    5.859e+05      4.328e+04      2.069e+04      1.187e+05     -3.557e+03
 yr_renovated  sqft_living15     sqft_lot15  is_renovated1   is_basement1
    3.319e+03      2.606e+01     -5.457e-01     -6.603e+06      7.603e+03
```



```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + floors +
    waterfront + view + condition + grade + yr_built + yr_renovated +
    sqft_living15 + sqft_lot15 + is_renovated, data = data1)

Coefficients:
  (Intercept)       bedrooms      bathrooms    sqft_living         floors
    6.211e+06     -3.951e+04      4.543e+04      1.632e+02      2.581e+04
  waterfront1           view      condition          grade       yr_built
    5.844e+05      4.388e+04      2.073e+04      1.187e+05     -3.577e+03
 yr_renovated  sqft_living15     sqft_lot15  is_renovated1
    3.338e+03      2.486e+01     -5.572e-01     -6.642e+06
```

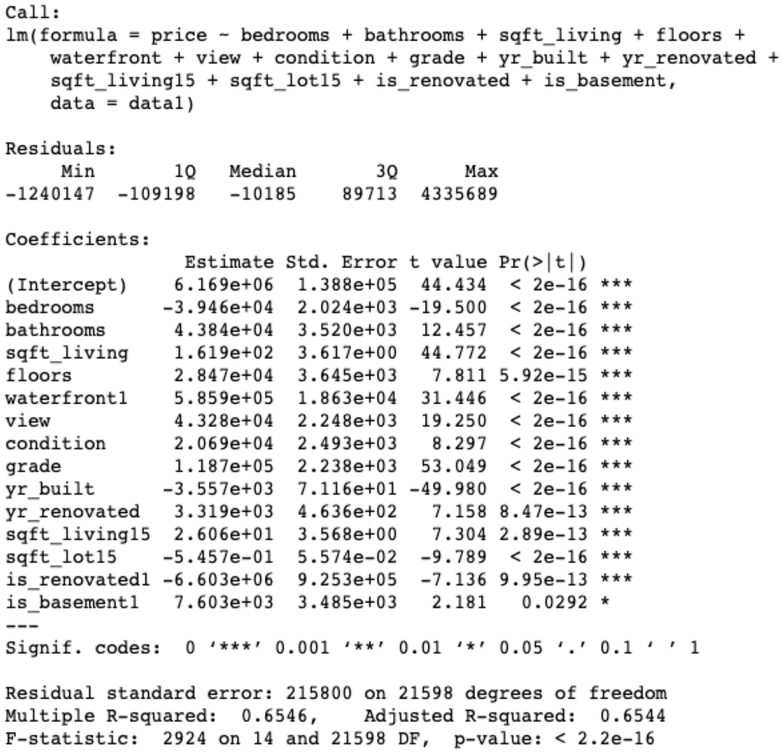
∗ Stepwise AIC & BIC



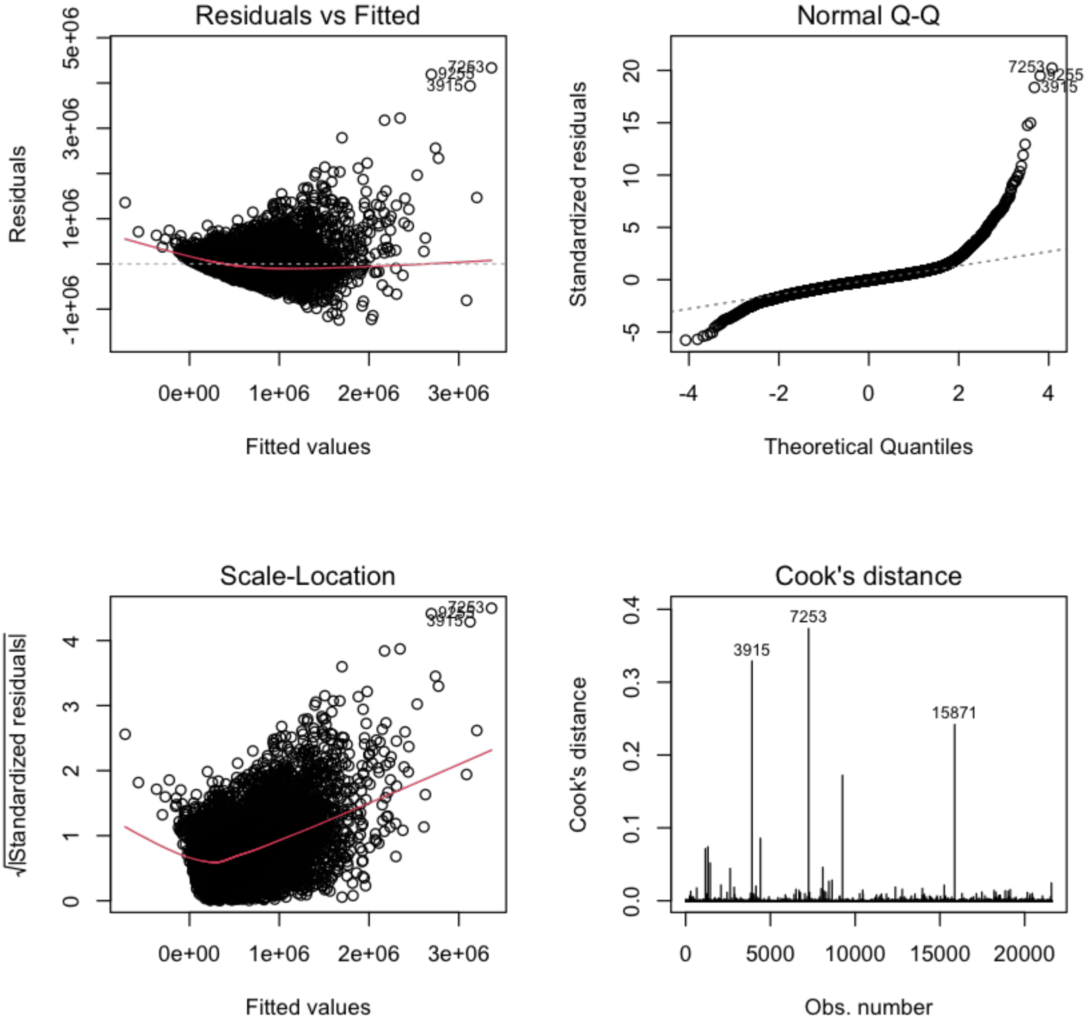
```
Call:
lm(formula = price ~ sqft_living + view + grade + yr_built +
    waterfront + bedrooms + bathrooms + sqft_lot15 + condition +
    floors + sqft_living15 + yr_renovated + is_renovated + is_basement,
    data = data1)

Coefficients:
  (Intercept)    sqft_living           view          grade       yr_built
    6.169e+06      1.619e+02      4.328e+04      1.187e+05     -3.557e+03
  waterfront1       bedrooms      bathrooms     sqft_lot15      condition
    5.859e+05     -3.946e+04      4.384e+04     -5.457e-01      2.069e+04
       floors  sqft_living15   yr_renovated  is_renovated1   is_basement1
    2.847e+04      2.606e+01      3.319e+03     -6.603e+06      7.603e+03


Call:
lm(formula = price ~ sqft_living + view + grade + yr_built +
    waterfront + bedrooms + bathrooms + sqft_lot15 + condition +
    floors + sqft_living15, data = data1)

Coefficients:
  (Intercept)    sqft_living           view          grade       yr_built
    6.331e+06      1.629e+02      4.395e+04      1.196e+05     -3.637e+03
  waterfront1       bedrooms      bathrooms     sqft_lot15      condition
    5.809e+05     -3.945e+04      4.771e+04     -5.573e-01      1.865e+04
       floors  sqft_living15
    2.513e+04      2.355e+01
```

From the above variable selection methods, we see that all the Forward, Backward & Stepwise AIC techniques yield 14 predictors, whereas the Forward & Stepwise BIC techniques yield 11 predictors, and the Backward BIC technique yields 13 predictors, respectively.

For my analysis, I have used the backward AIC technique to retain more number of predictors.



```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + floors +
    waterfront + view + condition + grade + yr_built + yr_renovated +
    sqft_living15 + sqft_lot15 + is_renovated + is_basement,
    data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1240147  -109198   -10185    89713  4335689

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.169e+06  1.388e+05  44.434  < 2e-16 ***
bedrooms      -3.946e+04  2.024e+03 -19.500  < 2e-16 ***
bathrooms      4.384e+04  3.520e+03  12.457  < 2e-16 ***
sqft_living    1.619e+02  3.617e+00  44.772  < 2e-16 ***
floors         2.847e+04  3.645e+03   7.811 5.92e-15 ***
waterfront1    5.859e+05  1.863e+04  31.446  < 2e-16 ***
view           4.328e+04  2.248e+03  19.250  < 2e-16 ***
condition      2.069e+04  2.493e+03   8.297  < 2e-16 ***
grade          1.187e+05  2.238e+03  53.049  < 2e-16 ***
yr_built      -3.557e+03  7.116e+01 -49.980  < 2e-16 ***
yr_renovated   3.319e+03  4.636e+02   7.158 8.47e-13 ***
sqft_living15  2.606e+01  3.568e+00   7.304 2.89e-13 ***
sqft_lot15    -5.457e-01  5.574e-02  -9.789  < 2e-16 ***
is_renovated1 -6.603e+06  9.253e+05  -7.136 9.95e-13 ***
is_basement1   7.603e+03  3.485e+03   2.181   0.0292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 215800 on 21598 degrees of freedom
Multiple R-squared:  0.6546,    Adjusted R-squared:  0.6544
F-statistic:  2924 on 14 and 21598 DF,  p-value: < 2.2e-16
```



Here, we see that the assumptions of MLR are still not satisfied based on the diagnostic plots.

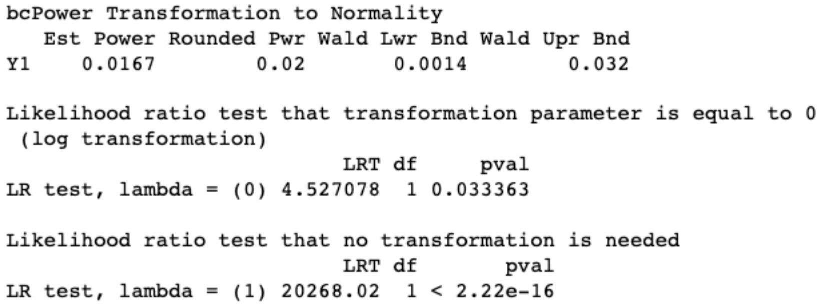
– Transformations

Let us use Power Transformation to decide whether the response and predictors need to be transformed.

∗ For predictors:

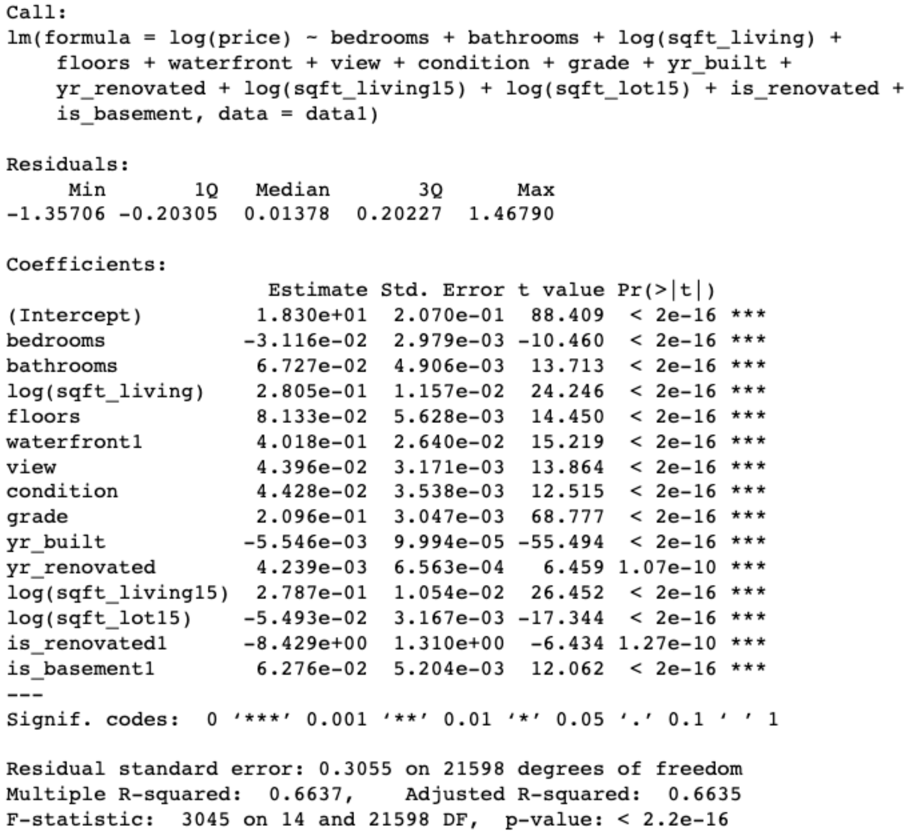
```
bcPower Transformations to Multinormality
              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
sqft_living      0.1333        0.13       0.1117       0.1549
sqft_living15   -0.0351       -0.04      -0.0645      -0.0057
sqft_lot15      -0.2065       -0.21      -0.2161      -0.1968

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                            LRT df       pval
LR test, lambda = (0 0 0) 2030.032  3 < 2.22e-16

Likelihood ratio test that no transformations are needed
                            LRT df       pval
LR test, lambda = (1 1 1) 77000.89  3 < 2.22e-16
```

- For response:



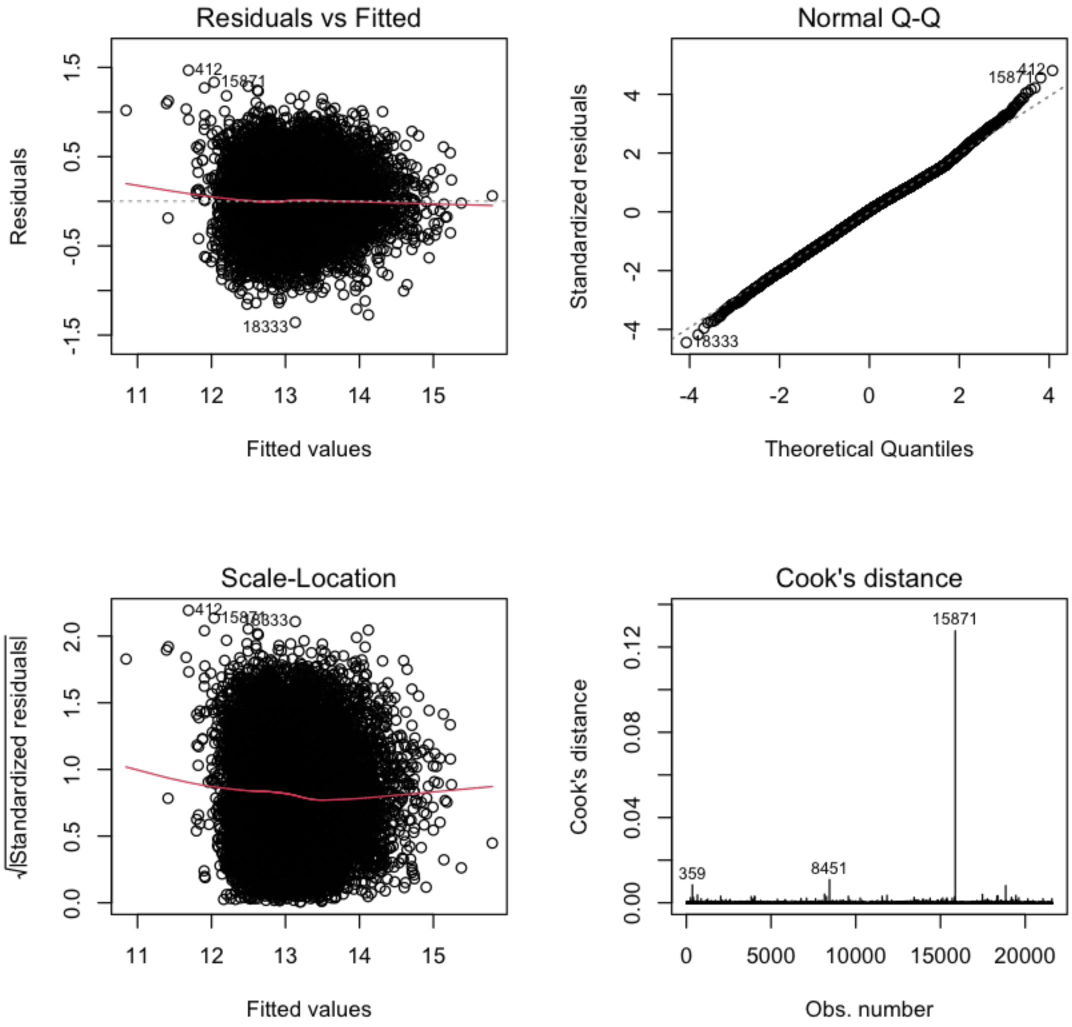
```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.0167        0.02       0.0014        0.032

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df      pval
LR test, lambda = (0) 4.527078  1 0.033363

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 20268.02  1 < 2.22e-16
```

- Summary after log transformations:



```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + log(sqft_living) +
    floors + waterfront + view + condition + grade + yr_built +
    yr_renovated + log(sqft_living15) + log(sqft_lot15) + is_renovated +
    is_basement, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35706 -0.20305  0.01378  0.20227  1.46790

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.830e+01  2.070e-01  88.409  < 2e-16 ***
bedrooms           -3.116e-02  2.979e-03 -10.460  < 2e-16 ***
bathrooms           6.727e-02  4.906e-03  13.713  < 2e-16 ***
log(sqft_living)    2.805e-01  1.157e-02  24.246  < 2e-16 ***
floors              8.133e-02  5.628e-03  14.450  < 2e-16 ***
waterfront1         4.018e-01  2.640e-02  15.219  < 2e-16 ***
view                4.396e-02  3.171e-03  13.864  < 2e-16 ***
condition           4.428e-02  3.538e-03  12.515  < 2e-16 ***
grade               2.096e-01  3.047e-03  68.777  < 2e-16 ***
yr_built           -5.546e-03  9.994e-05 -55.494  < 2e-16 ***
yr_renovated        4.239e-03  6.563e-04   6.459 1.07e-10 ***
log(sqft_living15)  2.787e-01  1.054e-02  26.452  < 2e-16 ***
log(sqft_lot15)    -5.493e-02  3.167e-03 -17.344  < 2e-16 ***
is_renovated1      -8.429e+00  1.310e+00  -6.434 1.27e-10 ***
is_basement1        6.276e-02  5.204e-03  12.062  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3055 on 21598 degrees of freedom
Multiple R-squared:  0.6637,    Adjusted R-squared:  0.6635
F-statistic:  3045 on 14 and 21598 DF,  p-value: < 2.2e-16
```

- Diagnostics after log transformations:

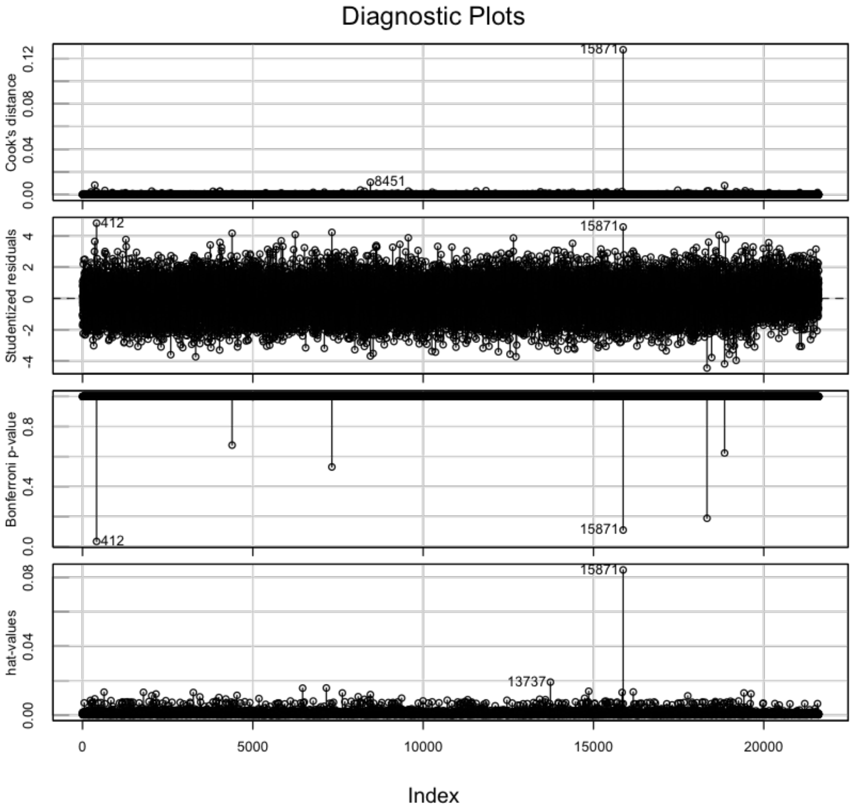


Here, the assumptions of MLR appear to be satisfied. However, from the Cook's distance plot, we see the presence of high influential points.

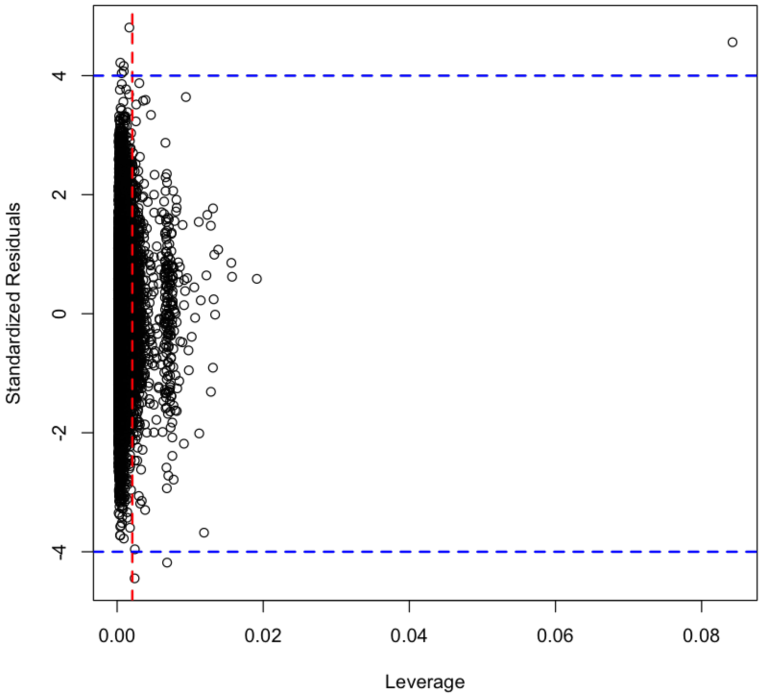
- Outliers

  As we can see from the above diagnostic plot (Cook's Distance plot), there seems to be a high influential point (15871). As it lies above 0.05 threshold, let us remove this -

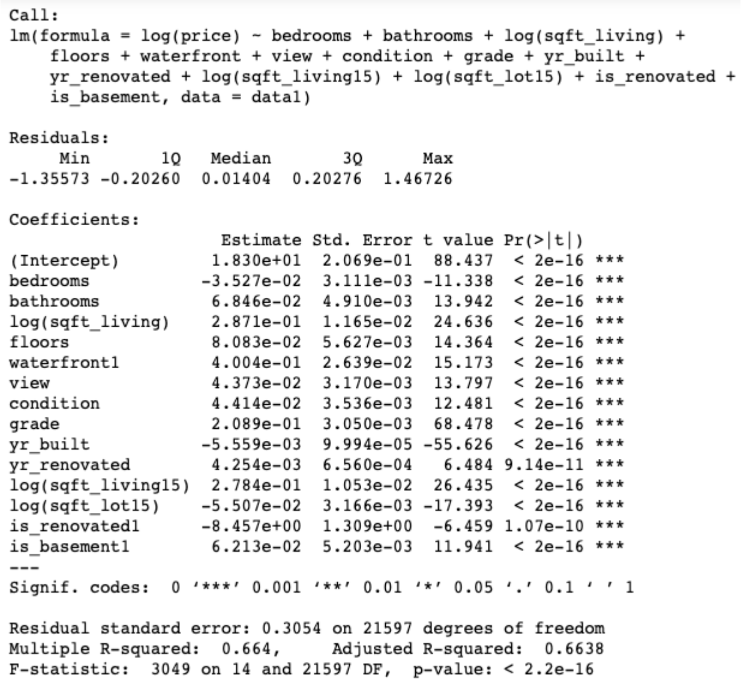
  * Influential Index Points:



∗ Standard Residual Vs. Leverage Plot:

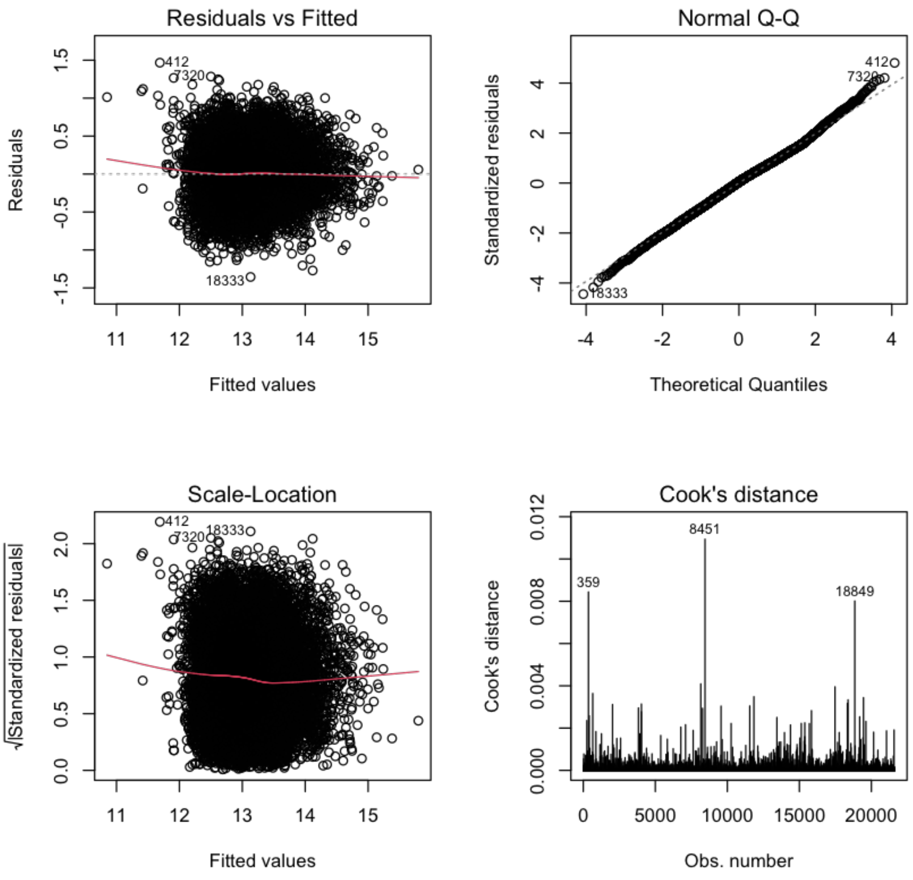


∗ Model Summary after removing high influential point (15871):

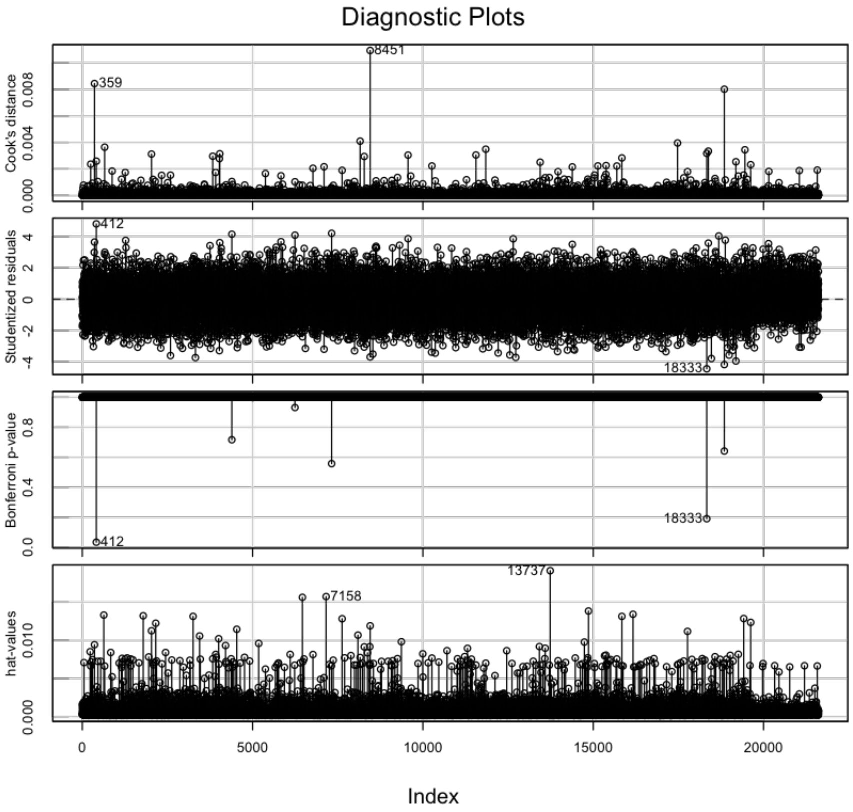


```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + log(sqft_living) +
    floors + waterfront + view + condition + grade + yr_built +
    yr_renovated + log(sqft_living15) + log(sqft_lot15) + is_renovated +
    is_basement, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.35573 -0.20260  0.01404  0.20276  1.46726

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.830e+01  2.069e-01  88.437  < 2e-16 ***
bedrooms           -3.527e-02  3.111e-03 -11.338  < 2e-16 ***
bathrooms           6.846e-02  4.910e-03  13.942  < 2e-16 ***
log(sqft_living)    2.871e-01  1.165e-02  24.636  < 2e-16 ***
floors              8.083e-02  5.627e-03  14.364  < 2e-16 ***
waterfront1         4.004e-01  2.639e-02  15.173  < 2e-16 ***
view                4.373e-02  3.170e-03  13.797  < 2e-16 ***
condition           4.414e-02  3.536e-03  12.481  < 2e-16 ***
grade               2.089e-01  3.050e-03  68.478  < 2e-16 ***
yr_built           -5.559e-03  9.994e-05 -55.626  < 2e-16 ***
yr_renovated        4.254e-03  6.560e-04   6.484 9.14e-11 ***
log(sqft_living15)  2.784e-01  1.053e-02  26.435  < 2e-16 ***
log(sqft_lot15)    -5.507e-02  3.166e-03 -17.393  < 2e-16 ***
is_renovated1      -8.457e+00  1.309e+00  -6.459 1.07e-10 ***
is_basement1        6.213e-02  5.203e-03  11.941  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3054 on 21597 degrees of freedom
Multiple R-squared:  0.664,     Adjusted R-squared:  0.6638
F-statistic:  3049 on 14 and 21597 DF,  p-value: < 2.2e-16
```

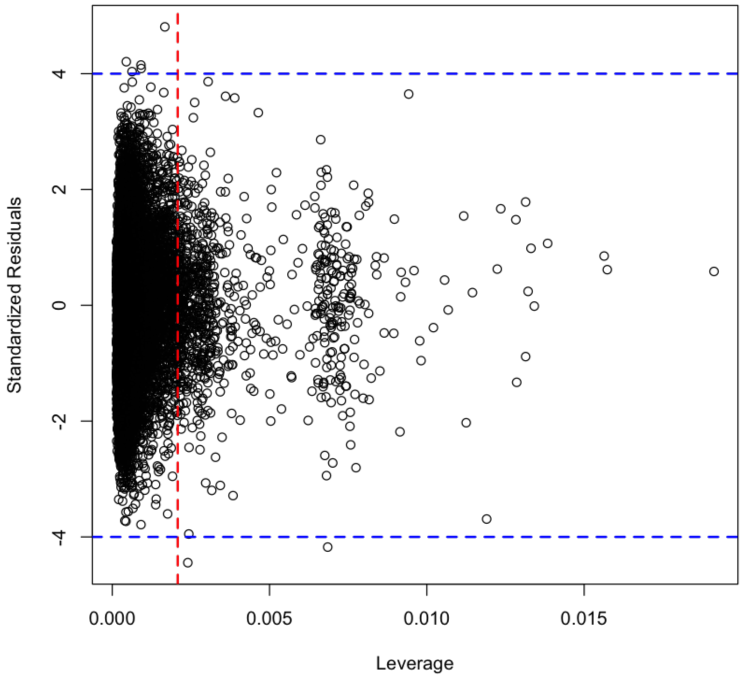
∗ Diagnostics after removing high influential point (15871):



∗ Influential Index Points:



∗ Standard Residual Vs. Leverage Plot:



Now, we see that the assumptions of MLR seem to be satisfied. We see the Normal Q-Q plot appears to be linear and the Residual Vs. Fitted plot does not appear to have constant variance. There also don't seem to be any bad leverage points. Therefore, we choose this as our final model to answer Research Question 1.

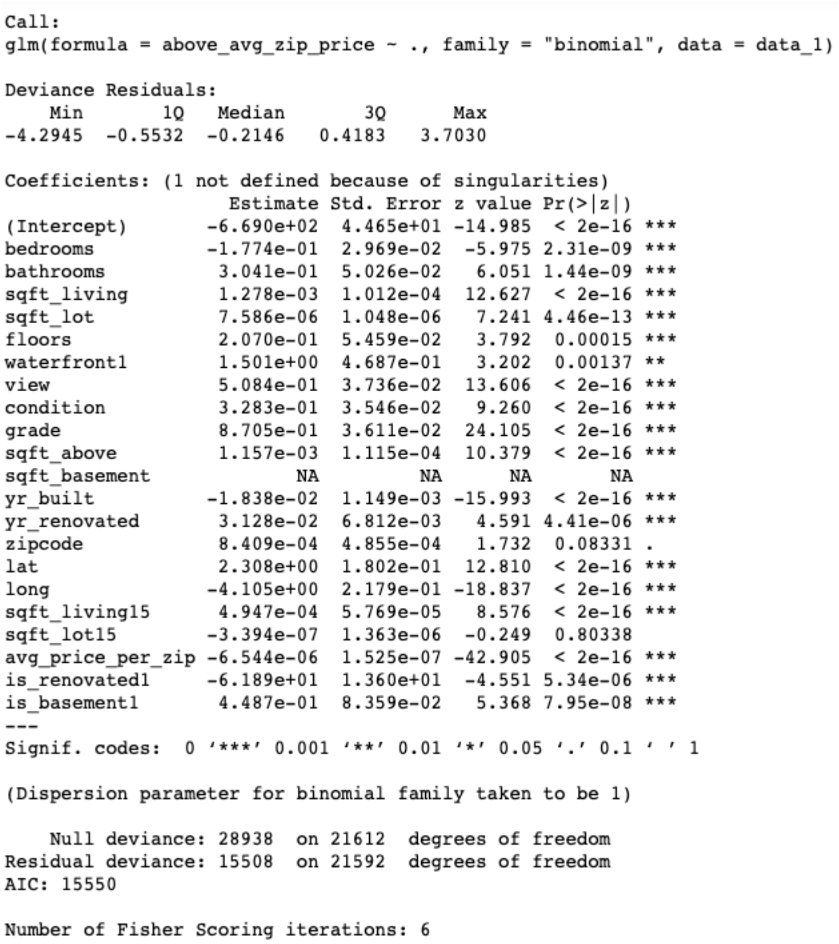
2.3.2 For Research Question 2

This being a logistic regression problem, we need to create a couple of additional variables:

- **avg_price_per_zip**: The average price of homes in a given zip code.
- **above_avg_zip_price**: A flag to indicate whether a given home's price is above or below the average price of homes in that zip code.

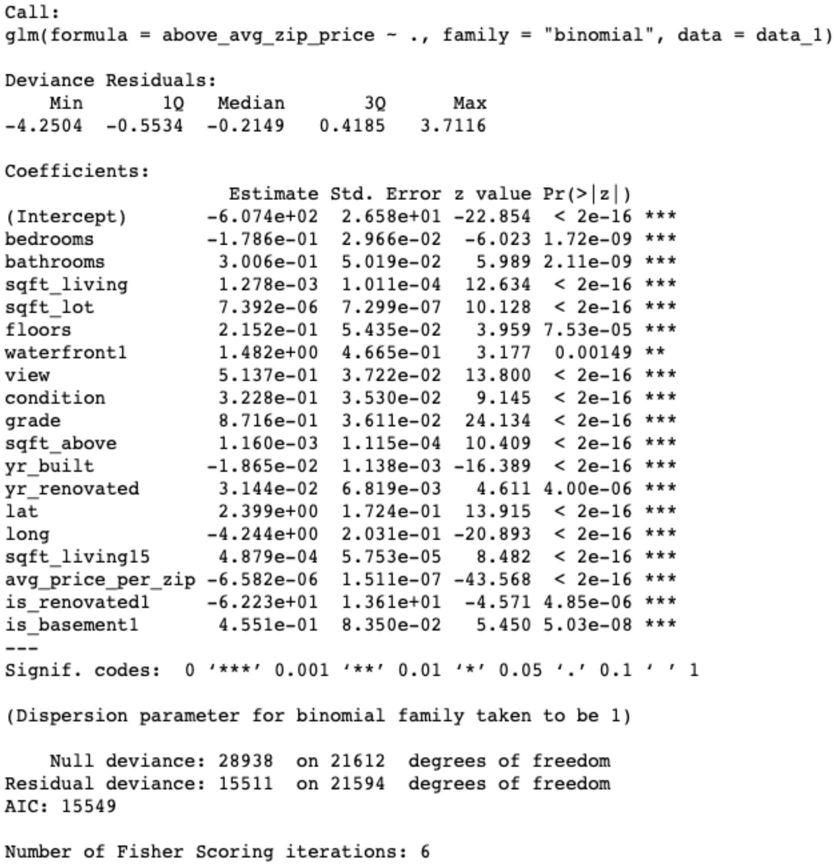
Next, since we are attempting to predict whether a house would sell above or below the asking price, we need to exclude the **price** feature from our data. Now our model is completely blind to the price of the home and will only make it's predictions based on the other features.

Initially, we attempt to fit a GLM model on all the features (except **id** & **date**) with **above_avg_zip_price** as the response.



```
Call:
glm(formula = above_avg_zip_price ~ ., family = "binomial", data = data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2945  -0.5532  -0.2146   0.4183   3.7030

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.690e+02  4.465e+01 -14.985  < 2e-16 ***
bedrooms          -1.774e-01  2.969e-02  -5.975 2.31e-09 ***
bathrooms          3.041e-01  5.026e-02   6.051 1.44e-09 ***
sqft_living        1.278e-03  1.012e-04  12.627  < 2e-16 ***
sqft_lot           7.586e-06  1.048e-06   7.241 4.46e-13 ***
floors             2.070e-01  5.459e-02   3.792  0.00015 ***
waterfront1        1.501e+00  4.687e-01   3.202  0.00137 **
view               5.084e-01  3.736e-02  13.606  < 2e-16 ***
condition          3.283e-01  3.546e-02   9.260  < 2e-16 ***
grade              8.705e-01  3.611e-02  24.105  < 2e-16 ***
sqft_above         1.157e-03  1.115e-04  10.379  < 2e-16 ***
sqft_basement             NA         NA      NA       NA
yr_built          -1.838e-02  1.149e-03 -15.993  < 2e-16 ***
yr_renovated       3.128e-02  6.812e-03   4.591 4.41e-06 ***
zipcode            8.409e-04  4.855e-04   1.732  0.08331 .
lat                2.308e+00  1.802e-01  12.810  < 2e-16 ***
long              -4.105e+00  2.179e-01 -18.837  < 2e-16 ***
sqft_living15      4.947e-04  5.769e-05   8.576  < 2e-16 ***
sqft_lot15        -3.394e-07  1.363e-06  -0.249  0.80338
avg_price_per_zip -6.544e-06  1.525e-07 -42.905  < 2e-16 ***
is_renovated1     -6.189e+01  1.360e+01  -4.551 5.34e-06 ***
is_basement1       4.487e-01  8.359e-02   5.368 7.95e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28938  on 21612  degrees of freedom
Residual deviance: 15508  on 21592  degrees of freedom
AIC: 15550

Number of Fisher Scoring iterations: 6
```

Here, we see that **sqft_basement**, **zipcode** & **sqft_lot15** are not significant. Hence, we exclude these from the model. Let us see the model summary with the final set of features:



```
Call:
glm(formula = above_avg_zip_price ~ ., family = "binomial", data = data_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-4.2504  -0.5534  -0.2149   0.4185   3.7116  

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -6.074e+02  2.658e+01 -22.854  < 2e-16 ***
bedrooms          -1.786e-01  2.966e-02  -6.023 1.72e-09 ***
bathrooms          3.006e-01  5.019e-02   5.989 2.11e-09 ***
sqft_living        1.278e-03  1.011e-04  12.634  < 2e-16 ***
sqft_lot           7.392e-06  7.299e-07  10.128  < 2e-16 ***
floors             2.152e-01  5.435e-02   3.959 7.53e-05 ***
waterfront1        1.482e+00  4.665e-01   3.177  0.00149 ** 
view               5.137e-01  3.722e-02  13.800  < 2e-16 ***
condition          3.228e-01  3.530e-02   9.145  < 2e-16 ***
grade              8.716e-01  3.611e-02  24.134  < 2e-16 ***
sqft_above         1.160e-03  1.115e-04  10.409  < 2e-16 ***
yr_built          -1.865e-02  1.138e-03 -16.389  < 2e-16 ***
yr_renovated       3.144e-02  6.819e-03   4.611 4.00e-06 ***
lat                2.399e+00  1.724e-01  13.915  < 2e-16 ***
long              -4.244e+00  2.031e-01 -20.893  < 2e-16 ***
sqft_living15      4.879e-04  5.753e-05   8.482  < 2e-16 ***
avg_price_per_zip -6.582e-06  1.511e-07 -43.568  < 2e-16 ***
is_renovated1     -6.223e+01  1.361e+01  -4.571 4.85e-06 ***
is_basement1       4.551e-01  8.350e-02   5.450 5.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28938  on 21612  degrees of freedom
Residual deviance: 15511  on 21594  degrees of freedom
AIC: 15549

Number of Fisher Scoring iterations: 6
```

To address the *reliability* clause in the research question, we split our data into train (70% of the data) and test sets (30% of the data). This ensures that the claimed model performance is on previously unseen data and hence more reliable. As the name suggests we fit the GLM model to our train dataset which consists of 15,129 records. The other 6484 records are hidden from the model during this time. Next, to verify the model's efficacy, we test the model on the held out data set. Using 0.5 as the threshold, we compute the probability outputs of the model and compare it with the actual target data. The model's performance is as follows -

– Confusion Matrix:

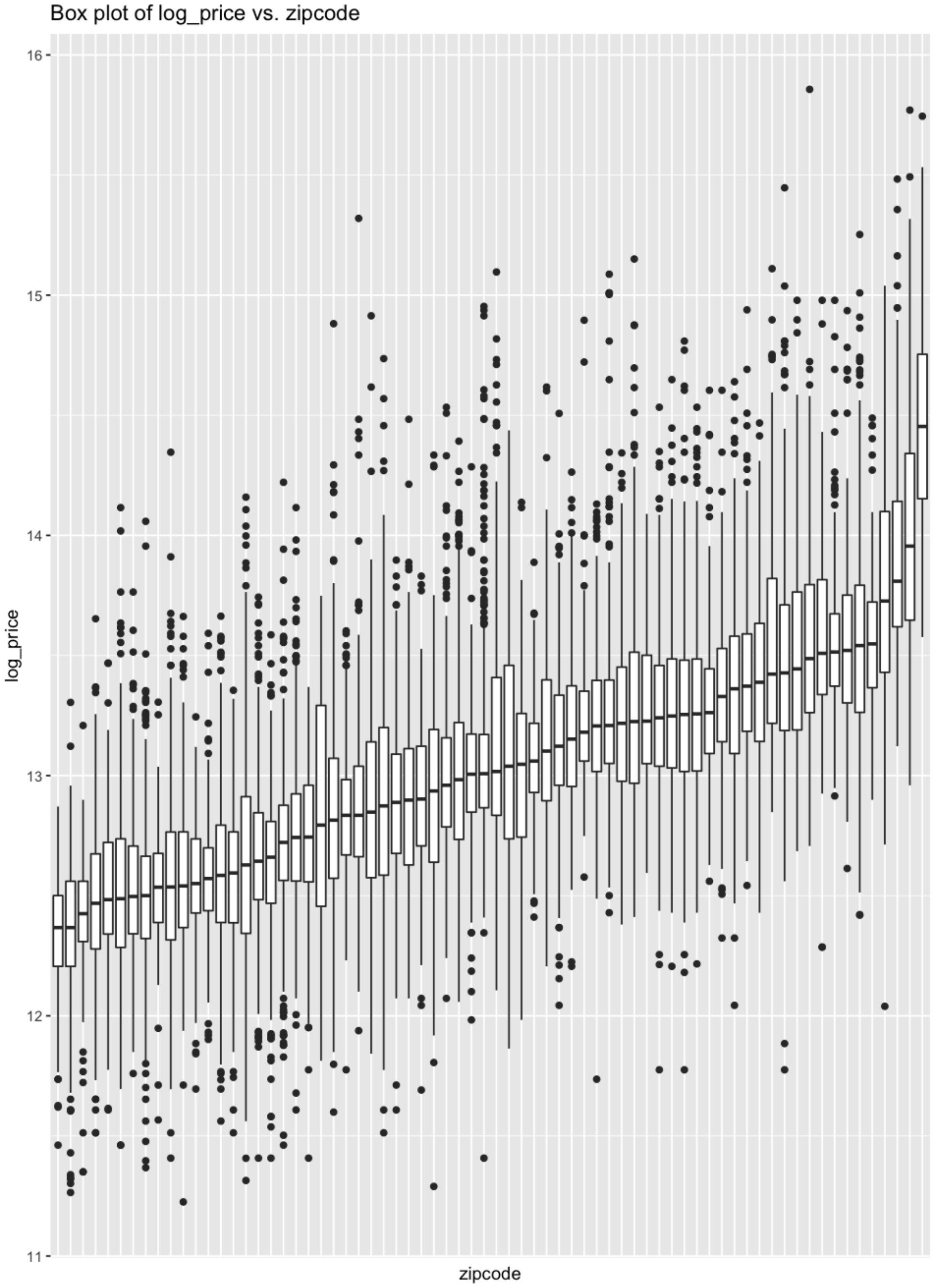
		Actual	
		0	1
Pred	0	3512	659
	1	403	1910

We choose this as our final model since it can accurately classify (Performance metrics discussed in *Interpretation* section) whether a home's value is above or below the average price of homes in that zip code. For the prediction component, we use the *Asking Price* as an input in place of the *avg_price_per_zip*. Since the model is trained on the avg. price per zip code of homes, this ensures the model is sensitive to variance between *Asking Price* & *avg_price_per_zip*, thus providing a rational classification.

2.3.3 For Research Question 3

By looking at the King County homes' distribution of zip codes and price, together with how a house is bought or sold by agents we make the following observations:

- The data set consists of 70 unique zip codes within King County.
- We see a clear difference between the price of homes in certain zip codes (Expensive neighborhoods Vs. Economical neighborhoods).
- Since the national average for agent commission for buying or selling a house is largely constant (3%), we can leverage our information about zip codes and price to maximize the agents' revenue for the number of homes sold.



Methodology

We employ the following steps to identify the high value zip codes:

- List individual zip codes and their respective median price of homes (for that individual zip code).
- Compare each row with the overall avg. median home price across King County.
- Add a new column reflecting the difference factor between each individual zip code's median home price and the overall avg. median home price across King County.
- Sort the table in descending order to identify the zip codes with the greatest positive variance from the overall avg. median home price across King County and shortlist the top 10 zip codes.

From our analysis, we see that by analyzing historical data of median home prices grouped by zip code, we can identify more expensive neighbourhoods. As we can see, we've identified the following zip codes as a part of the top 10:

98039 (Medina, which is the most expensive city of the Seattle metropolitan area), 98004, 98040, 98112, 98005, 98006, 98119, 98075, 98109 & 98102.

From an agent's perspective, since the amount of time and effort in selling a house is the same, focusing their efforts in zip codes have significantly higher median home prices would yield higher revenue. By carrying out business in the top 10 zip codes, as against random zip codes, one can expect to generate 1.87 times the revenue. (Revenue gain discussed in the *Interpretation* section)

2.4 Interpretation

2.4.1 Interpretation for Research Question 1

The equation for the MLR model would be like so:

$$log(price) = \beta_0 + \beta_1 bedrooms + \beta_2 bathrooms + \beta_3 log(sqft\_living) + \beta_4 floors + \beta_5 waterfront + \beta_6 view + \beta_7 condition + \beta_8 grade + \beta_9 yr\_built + \beta_{10} yr\_renovated + \beta_{11} log(sqft\_living15) + \beta_{12} log(sqft\_lot15) + \beta_{13} is\_renovated + \beta_{14} is\_basement + e$$

As seen from the equation above we can interpret the relationship between *log(price)* and the predictors in a linear fashion. The final set of features selected for the model were found to satisfy the assumptions of Multiple Linear Regression and also have shown to be important features for determining the response. Mapping the first research question to a real world scenario, we can predict the *log(price)* of a home given its set of predictors using this model and take the inverse log of the prediction to obtain the expected home selling price.

2.4.2 Interpretation for Research Question 2

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = 0.83$$

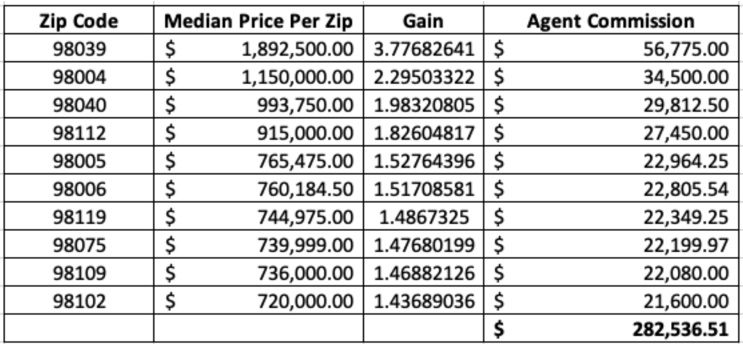
$$\text{Precision} = \frac{TP}{TP+FP} = 0.82$$

$$\text{Recall} = \frac{TP}{TP+FN} = 0.74$$

For our research question, since we are interested in seeing for how many of the houses can we correctly identify the price being above or below the avg. price of homes in it's zip code, **Accuracy** would be the most relevant metric.

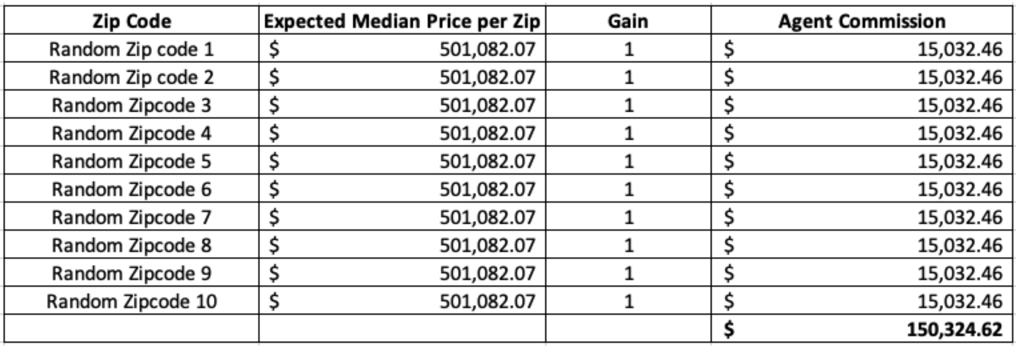
The model yields an Accuracy of over 83%. Since Accuracy is simply a ratio of correctly predicted observation to the total observations, we can translate it as the model's the ability to identify 83% of the homes's price being above or below the avg. price of homes in a given zip code.

2.4.3 Interpretation for Research Question 3



Zip Code	Median Price Per Zip	Gain	Agent Commission
98039	$ 1,892,500.00	3.77682641	$ 56,775.00
98004	$ 1,150,000.00	2.29503322	$ 34,500.00
98040	$ 993,750.00	1.98320805	$ 29,812.50
98112	$ 915,000.00	1.82604817	$ 27,450.00
98005	$ 765,475.00	1.52764396	$ 22,964.25
98006	$ 760,184.50	1.51708581	$ 22,805.54
98119	$ 744,975.00	1.4867325	$ 22,349.25
98075	$ 739,999.00	1.47680199	$ 22,199.97
98109	$ 736,000.00	1.46882126	$ 22,080.00
98102	$ 720,000.00	1.43689036	$ 21,600.00
			$ 282,536.51

The first table shows the expected revenue in agent commissions by selling a home in the suggested top 10 zip codes.



Zip Code	Expected Median Price per Zip	Gain	Agent Commission
Random Zip code 1	$ 501,082.07	1	$ 15,032.46
Random Zip code 2	$ 501,082.07	1	$ 15,032.46
Random Zipcode 3	$ 501,082.07	1	$ 15,032.46
Random Zipcode 4	$ 501,082.07	1	$ 15,032.46
Random Zipcode 5	$ 501,082.07	1	$ 15,032.46
Random Zipcode 6	$ 501,082.07	1	$ 15,032.46
Random Zipcode 7	$ 501,082.07	1	$ 15,032.46
Random Zipcode 8	$ 501,082.07	1	$ 15,032.46
Random Zipcode 9	$ 501,082.07	1	$ 15,032.46
Random Zipcode 10	$ 501,082.07	1	$ 15,032.46
			$ 150,324.62

The second table shows the expected revenue in agent commissions by selling a home in a random zip code of King County.

From the above tables:

- Expected revenue from top 10 zip codes: $282.5K
- Expected revenue from top Random zip codes: $150.3K
- Expected gain in revenue for agent: 1.87x

Conclusion

3.0.1 For Research Question 1

Using our results from Experiment 1, we conclude that we are not only able to establish a linear relationship between the the various features of housing data and the corresponding price, but also quantify this relationship with the Adj. R^2 value of 0.66 obtained from the resulting models. The Multiple Linear Regression model is able to explain over 66% of a home's selling prices based on it's features.

These features and the resulting model would serve as a base for future work such as inclusion of additional features, model optimization and application of more advanced models.

3.0.2 For Research Question 2

Using the results from Experiment 2, we conclude that we can accurately classify (with over 83% accuracy) whether a home's price is above or below the average price of homes in that zip code using a Logistic Regression model. The information from this model could be of assistance to aspiring home owners who are looking to place their bids on homes that are for sale.

3.0.3 For Research Question 3

From our experiment, we conclude that by analyzing historical data of median home prices grouped by zip code, we can identify more expensive neighbourhoods and also the less expensive ones. From an agent's perspective, since the amount of time and effort in selling a house is the same, focusing their efforts in zip codes have significantly higher median home prices would yield higher revenue. Leveraging the data from our analysis, we can not only provide a list of most desirable zip codes to maximize their revenue, but also a list of zip codes to avoid, which we know would yield significantly less revenue than others. This would be beneficial to agents who are looking to make data driven decisions in their strive to compete in the real estate market.

References

[1] STAT-632 Lecture Slides by Prof. Jiyoun Myung, 2021.

[2] *The LaTeX Companion* by Addison-Wesley, Reading, Massachusetts, 1993

[3] A Modern Approach to Regression with R by Simon J. Sheather

[4] R for Data Science by Hadley Wickham

Acknowledgement

I would like to express my heartfelt gratitude for Prof. Jiyoun Myung for her continued guidance and support. Learning remotely in 2020-2021 is tough, thank you for making it slightly easier for all of us. I would also like to thank my fellow classmates with whom I've had the opportunity to have many constructive conversations through the course of this semester. A special mention to all the thought provoking questioners during my demo, who helped me improve my project and present a more refined version in the report.

I feel obliged to offer my token of appreciation for Prof. Joshua Kerr, my advisor, Prof. Eric Suess, who has been a great mentor as a part of the Data Science Workgroup, who have been extremely helpful and have made me a better graduate student.

Finally, I would like to thank my family who have been extremely supportive and encouraging of my pursuits and enabled me to excel in my studies.

Appendices

Appendix 1: R Code

Link to Github Repo

(For print version of the report, please scan above QR code to view the Github code repository)

Appendix 2

- **Overall F test**:
  $H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_{17} = 0$
  H_A : At least one $\beta_j \neq 0, j = 1,2,3,4,...17$

```
Analysis of Variance Table

Model 1: price ~ 1
Model 2: price ~ bedrooms + bathrooms + sqft_living + sqft_lot + floors +
    waterfront + view + condition + grade + sqft_above + sqft_basement +
    yr_built + yr_renovated + sqft_living15 + sqft_lot15 + is_renovated +
    is_basement
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1  21435 2.8920e+15
2  21419 1.0003e+15 16 1.8917e+15 2531.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For the F-test we are comparing the full model, with all the predictors, to the null model, with no predictors. Since the p-value is less than 0.001, we reject the null hypothesis that $H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_{17} = 0$. Thus, we conclude, that at least one predictor is associated with Price.