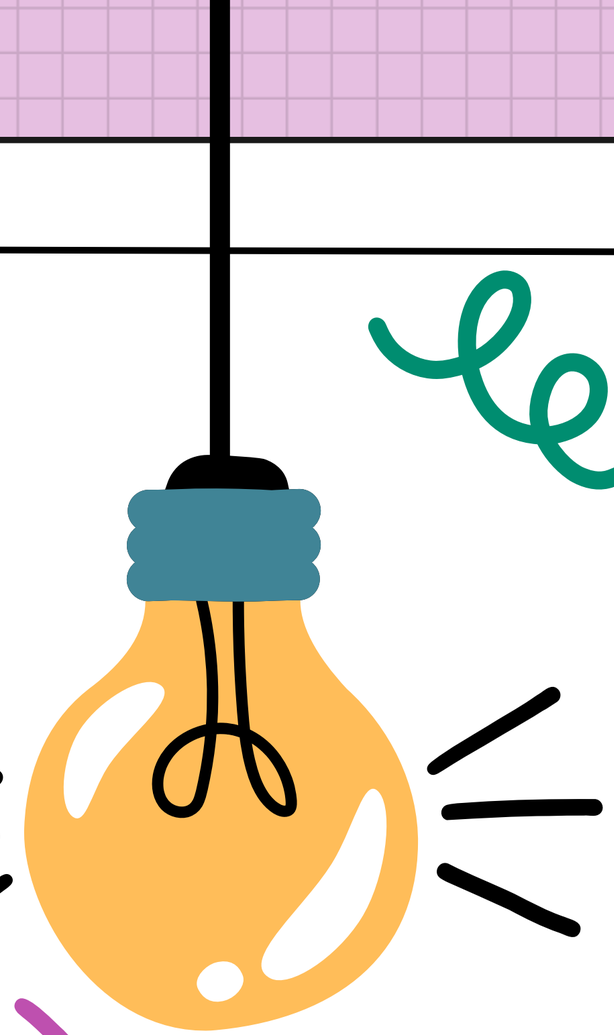
MODEL LINEAR

FINAL PROJECT



Group E

Analisis Regresi Terkait Harga Penjualan Rumah Menggunakan Model Regresi Linear Berganda

presented by :

Aulia Rahma Salsabila (2206032311)
Alma Ruth Junita Hutasoit (2206051903)
Natasya Ramadhani Dwi Putri (2206811751)
Razaan Hikmal Fasya (2206025464)
Tshabitha Aqnia Putri (2206051821)

PENDAHULUAN

Rincian Data

Akan digunakan dataset House Price. Dataset ini memuat harga rumah beserta spesifikasinya masing masing. Dataset ini memuat 4600 baris data dengan 18 kolom. Deskripsi kolom-kolom tersebut adalah sebagia berikut:

Date (String)	: Tanggal diuploadnya data	Sqft_above (Numerik)	: Luas lantai atas
Bedrooms (Numerik)	: Banyak kamar tidur	Sqft_basement (Numerik)	: Luas basement
Bathrooms (Numerik)	: Banyak kamar mandi	Yr_built (Numerik)	: Tahun dibuat
Sqft_Living (Numerik)	: Luas rumah	Yr_renovated (Numerik)	: Tahun terakhir rumah direnovasi
Sqft_lot (Numerik)	: Luas tanah	Street (String)	: Jalan rumah
Floors (Numerik)	: Jumlah lantai	City (String)	: Kota rumah
Waterfront (Categorical)	: Daerah dengan air	StateZip (String)	: Kode pos rumah
View (Numerik)	: Pemandangan dari rumah	Country (String)	: Negara rumah
Condition (Ordinal)	: Kondisi rumah	Price (Numerik)	: Harga rumah

PRE-PROCESSING DAN ANALISIS DESKRIPTIF

- Pemilihan kolom

Akan dilakukan penghapusan kolom berikut :

1. Date: Tanggal upload tidak akan dibutuhkan
2. Street: Nama jalan rumah terlalu beragam sehingga tidak dapat dijadikan categorical variable
3. City: Nama kota rumah terlalu beragam sehingga tidak dapat dijadikan categorical variable
4. StateZip: Kode pos rumah tidak akan dibutuhkan
5. Country: Hanya terdapat 1 nilai pada variable rumah sehingga dapat dihapus

- Missing Values

```
> data %>%
+   summarise(count = sum(is.na(data)))
  count
1     0
```

PRE-PROCESSING DAN ANALISIS DESKRIPTIF

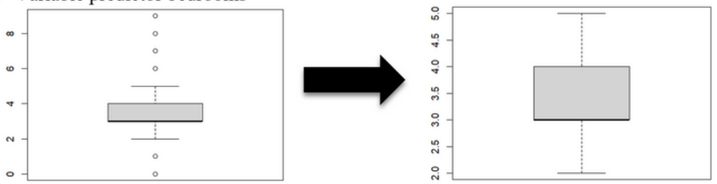
- **Outliers**

```
#Outliers
boxplot(data$bedrooms)
data = droplevels(data[-which(data$bedrooms>5),])
data = droplevels(data[-which(data$bedrooms<2),])
boxplot(data$bedrooms)

boxplot(data$bathrooms)
data = droplevels(data[-which(data$bathrooms>3.5),])
boxplot(data$bathrooms)

boxplot(data$sqft_living)
data = droplevels(data[-which(data$sqft_living>4050),])
boxplot(data$sqft_living)

boxplot(data$sqft_lot)
data = droplevels(data[-which(data$sqft_lot>15500),])
boxplot(data$sqft_lot)

boxplot(data$sqft_above)
data = droplevels(data[-which(data$sqft_above>3250),])
boxplot(data$sqft_above)

boxplot(data$sqft_basement)
data = droplevels(data[-which(data$sqft_basement>1350),])
boxplot(data$sqft_basement)
```

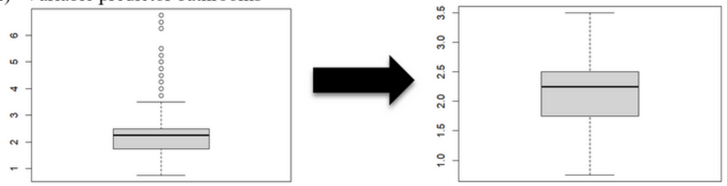
PRE-PROCESSING DAN ANALISIS DESKRIPTIF

Hasil dari penghapusan outliers:

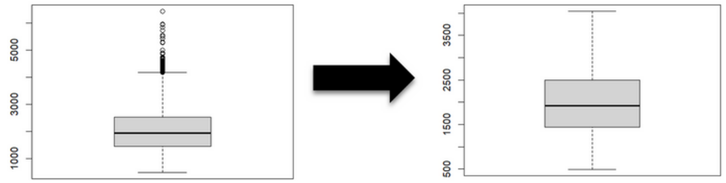
1) Variable predictor bedrooms



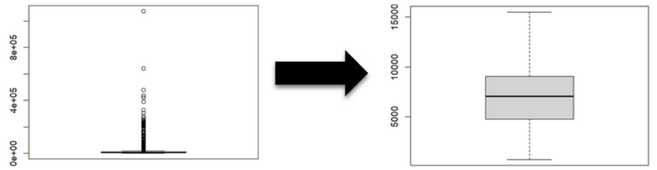
2) Variable predictor bathrooms



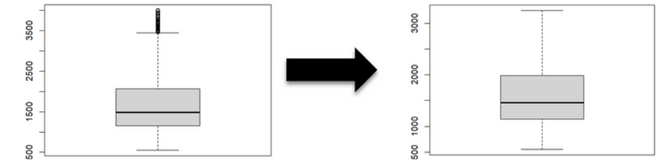
3) Variable predictor sqft_living



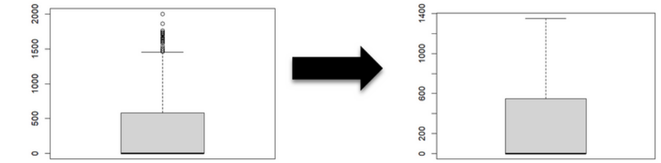
4) Variable predictor sqft_lot



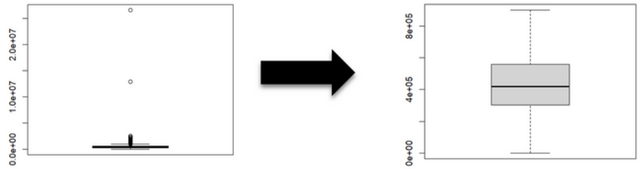
5) Variable predictor sqft_above



6) Variable predictor sqft_basement

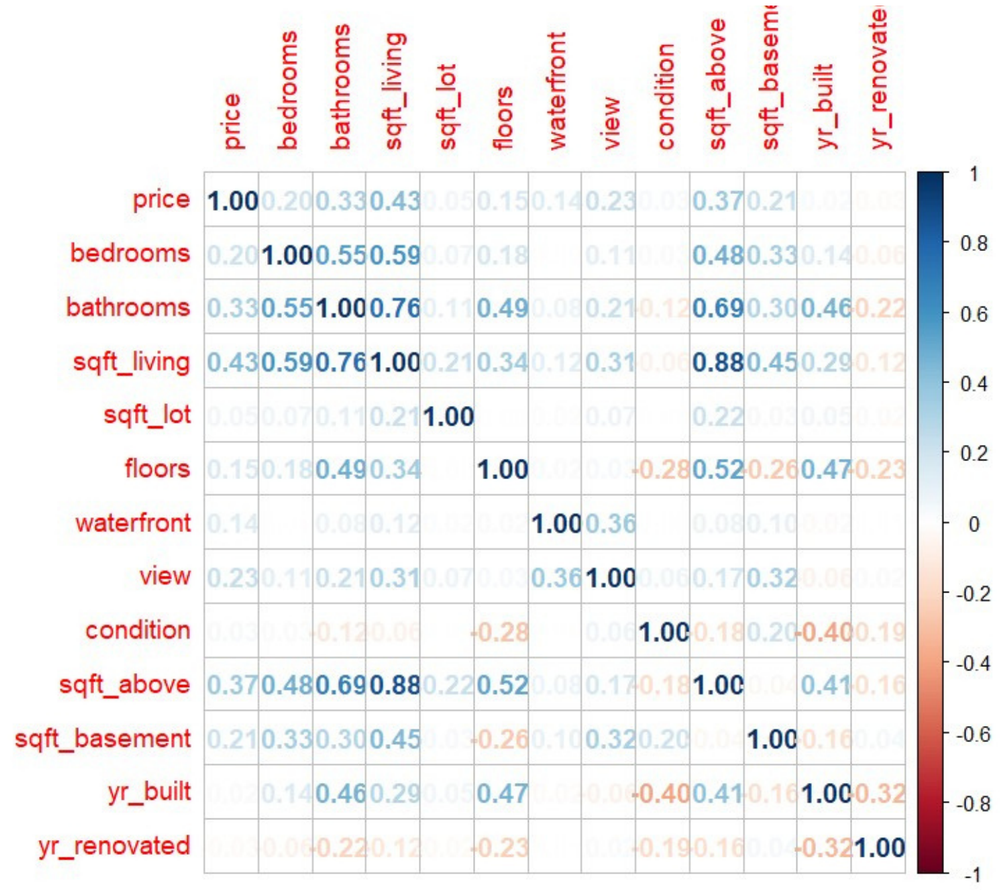


7) Variable respon price



PRE-PROCESSING DAN ANALISIS DESKRIPTIF

Analisis Deskriptif

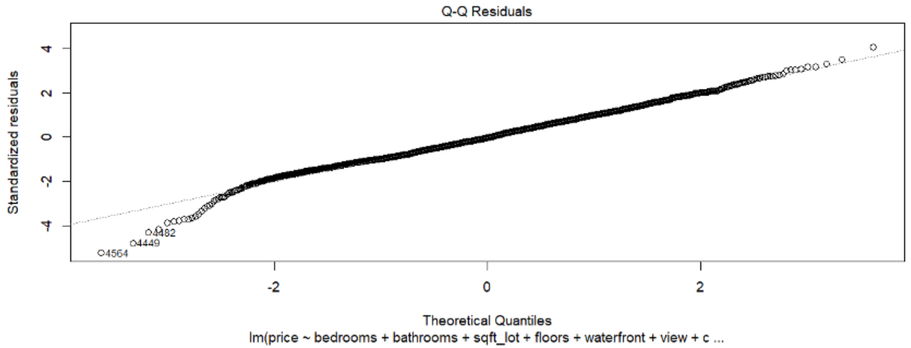


PEMODELAN

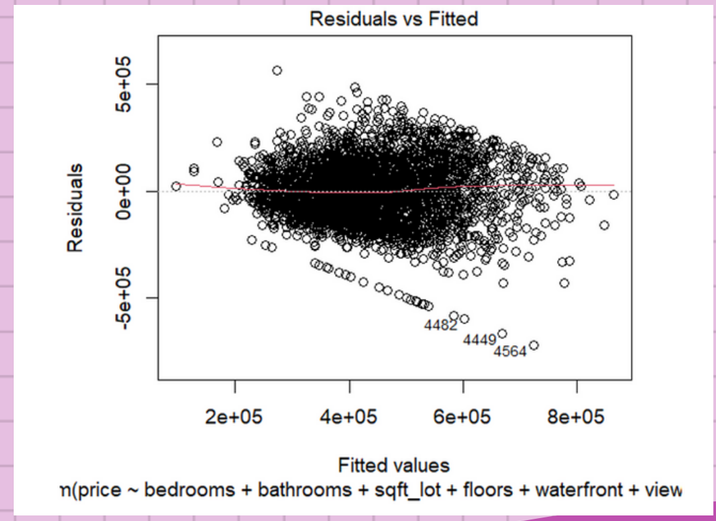
MODEL PERTAMA

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11}$$

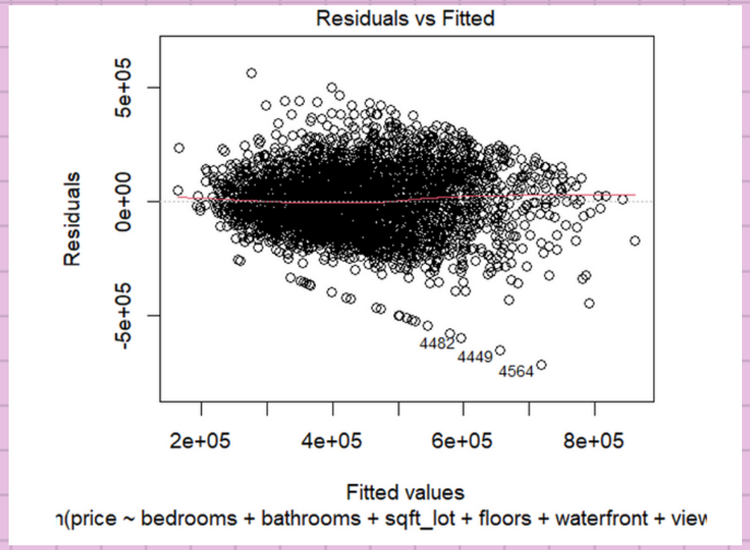
```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
    waterfront + view + condition + sqft_above + sqft_basement +
    yr_built + yr_renovated, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-723811  -95002   -4560   93867  565740

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.623e+06  2.466e+05  14.692  < 2e-16 ***
bedrooms     -3.740e+04  4.027e+03  -9.288  < 2e-16 ***
bathrooms     4.134e+04  6.397e+03   6.461 1.19e-10 ***
sqft_lot     -7.162e+00  9.684e-01  -7.396 1.76e-13 ***
floors        2.953e+04  7.195e+03   4.104 4.16e-05 ***
waterfront    4.470e+04  1.011e+05   0.442    0.659
view          2.609e+04  5.043e+03   5.174 2.43e-07 ***
condition2   -5.736e+04  1.053e+05  -0.545    0.586
condition3    6.966e+04  9.924e+04   0.702    0.483
condition4    8.269e+04  9.911e+04   0.834    0.404
condition5    1.038e+05  9.927e+04   1.045    0.296
sqft_above    1.826e+02  6.760e+00  27.017  < 2e-16 ***
sqft_basement 1.607e+02  9.001e+00  17.853  < 2e-16 ***
yr_built     -1.800e+03  1.197e+02 -15.043  < 2e-16 ***
yr_renovated  2.339e+00  2.931e+00   0.798    0.425
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 139700 on 3349 degrees of freedom
Multiple R-squared:  0.3894,    Adjusted R-squared:  0.3868
F-statistic: 152.5 on 14 and 3349 DF,  p-value: < 2.2e-16
```




```
> vif(model_1)
                   GVIF Df GVIF^(1/(2*Df))
bedrooms       1.657165  1        1.287309
bathrooms      2.945910  1        1.716365
sqft_lot       1.614989  1        1.270822
floors         2.639248  1        1.624576
waterfront     1.047234  1        1.023345
view           1.087780  1        1.042967
condition      1.710180  4        1.069376
sqft_above     2.757410  1        1.660545
sqft_basement  1.908012  1        1.381308
yr_built       2.259076  1        1.503022
yr_renovated   1.428604  1        1.195242
```

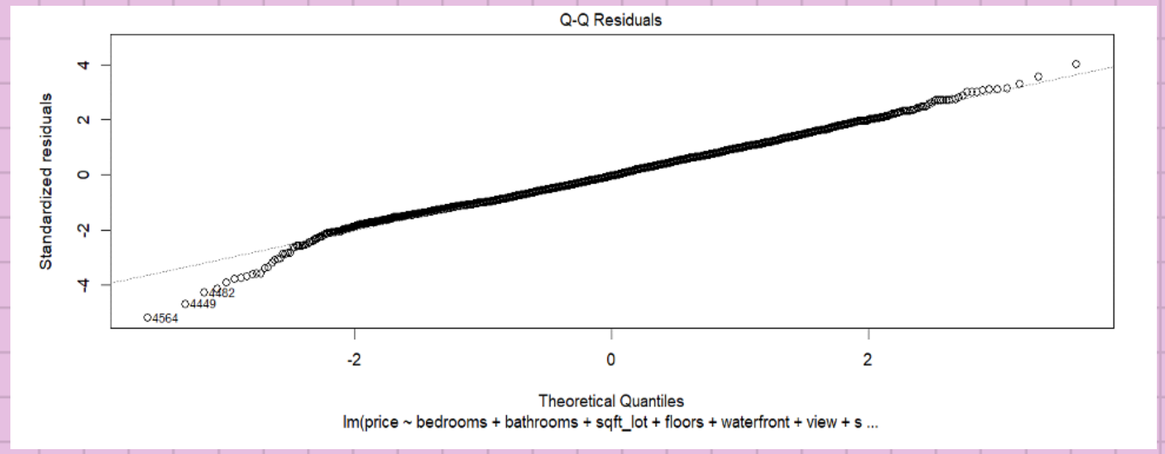


MODEL KEDUA



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}$$

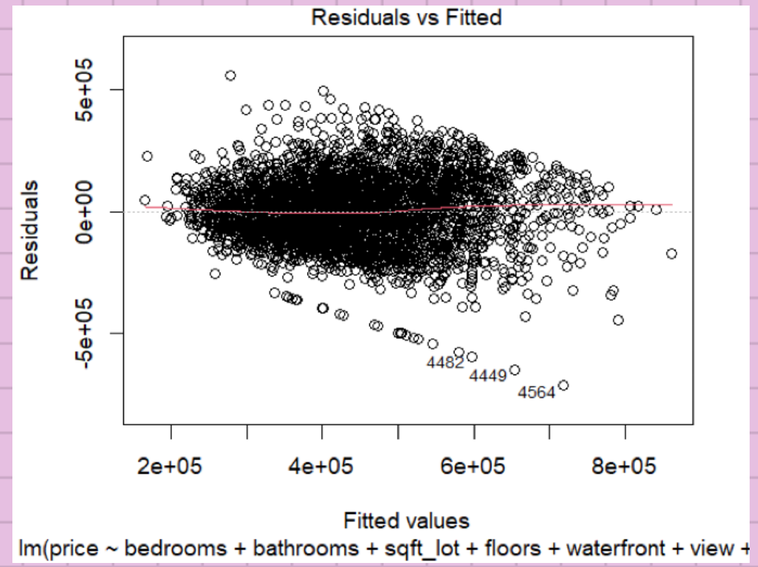
```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
    waterfront + view + sqft_above + sqft_basement + yr_built +
    yr_renovated, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-717200  -96533   -3409   93128  561556

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.012e+06  2.086e+05  19.229  < 2e-16 ***
bedrooms     -3.708e+04  4.036e+03  -9.187  < 2e-16 ***
bathrooms     4.370e+04  6.380e+03   6.850 8.77e-12 ***
sqft_lot     -6.956e+00  9.608e-01  -7.240 5.55e-13 ***
floors        2.764e+04  7.183e+03   3.848 0.000121 ***
waterfront    5.642e+04  1.014e+05   0.556 0.578162
view          2.586e+04  5.058e+03   5.113 3.35e-07 ***
sqft_above    1.814e+02  6.768e+00  26.804  < 2e-16 ***
sqft_basement 1.626e+02  9.021e+00  18.025  < 2e-16 ***
yr_built     -1.959e+03  1.086e+02 -18.046  < 2e-16 ***
yr_renovated -9.341e-01  2.637e+00  -0.354 0.723164
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140200 on 3353 degrees of freedom
Multiple R-squared:  0.3843,    Adjusted R-squared:  0.3824
F-statistic: 209.3 on 10 and 3353 DF,  p-value: < 2.2e-16
```

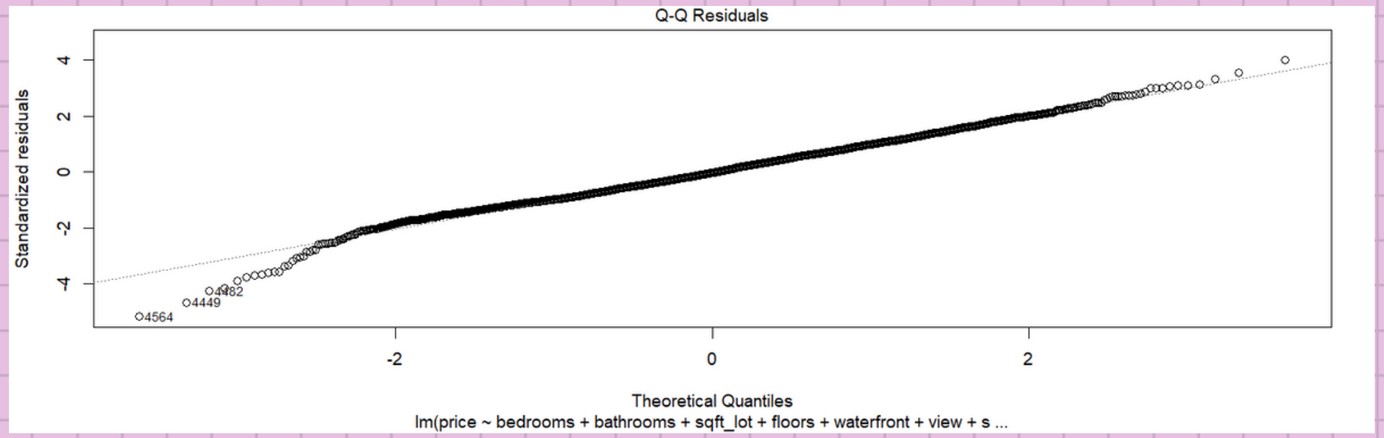



```
> vif(model_2)
    bedrooms    bathrooms     sqft_lot       floors   waterfront         view
    1.652931     2.908853     1.578371     2.611843     1.046174     1.086661
  sqft_above sqft_basement     yr_built yr_renovated
    2.744805     1.903051     1.846063     1.148146
```

MODEL KETIGA

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9$$



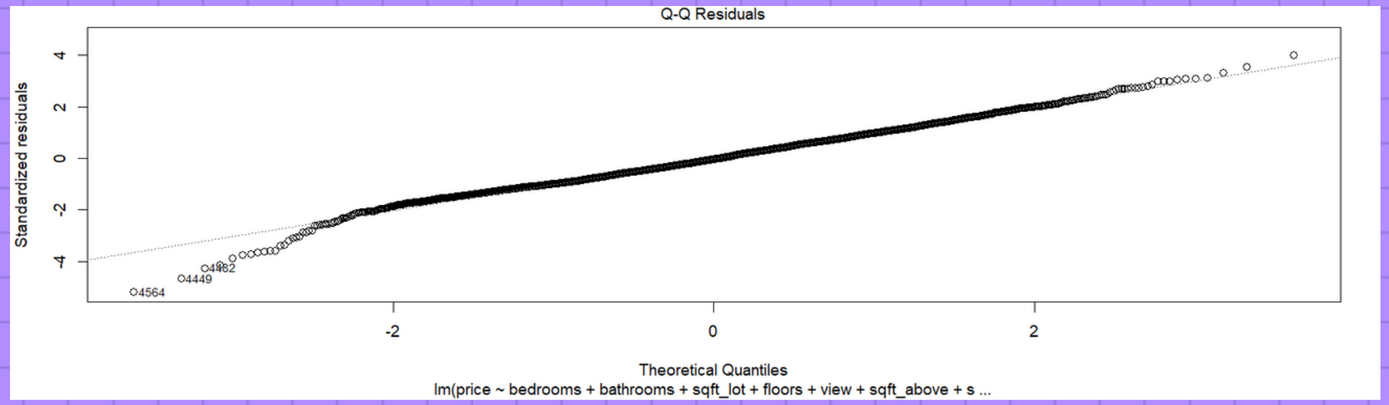



```
> vif(model_3)
    bedrooms    bathrooms     sqft_lot       floors   waterfront         view
    1.652450     2.884413     1.576125     2.605645     1.045278     1.085944
    yr_built   sqft_above sqft_basement
    1.795456     2.744477     1.903050
```

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
    waterfront + view + sqft_above + sqft_basement + yr_built,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-718044  -97019   -3595   93437  560877

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.998e+06  2.050e+05  19.504  < 2e-16 ***
bedrooms     -3.711e+04  4.035e+03  -9.195  < 2e-16 ***
bathrooms     4.390e+04  6.352e+03   6.912 5.69e-12 ***
sqft_lot     -6.968e+00  9.600e-01  -7.259 4.82e-13 ***
floors        2.776e+04  7.173e+03   3.870 0.000111 ***
waterfront    5.747e+04  1.014e+05   0.567 0.570878
view          2.582e+04  5.056e+03   5.106 3.47e-07 ***
sqft_above    1.814e+02  6.767e+00  26.805  < 2e-16 ***
sqft_basement 1.626e+02  9.020e+00  18.028  < 2e-16 ***
yr_built     -1.953e+03  1.070e+02 -18.241  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140200 on 3354 degrees of freedom
Multiple R-squared:  0.3843,    Adjusted R-squared:  0.3826
F-statistic: 232.6 on 9 and 3354 DF,  p-value: < 2.2e-16
```

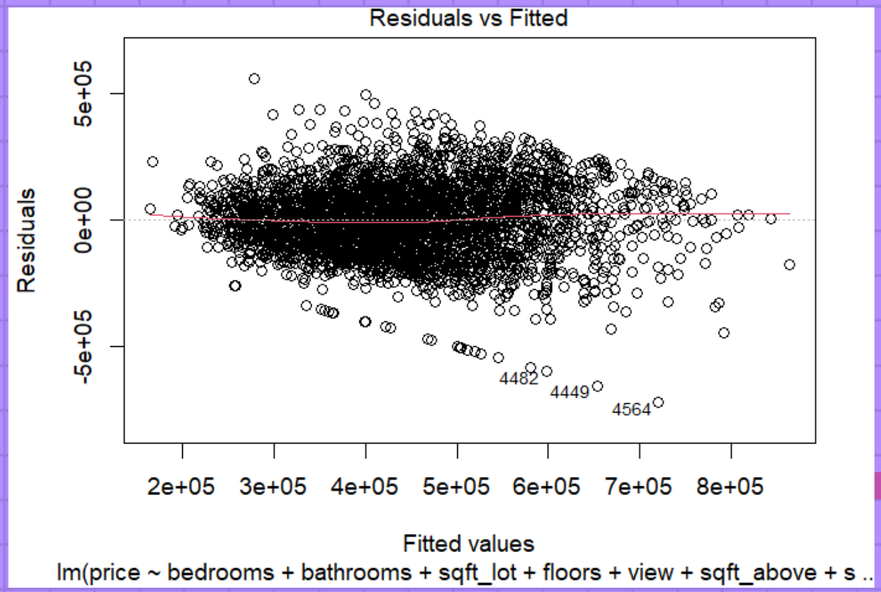
MODEL KEEMPAT

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$



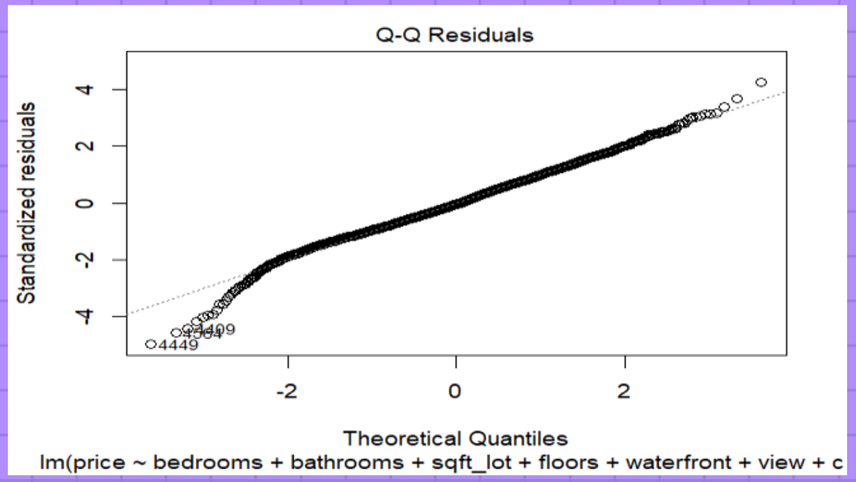

```
> vif(model_4)
    bedrooms    bathrooms      sqft_lot       floors
    1.649972     2.883137     1.568961     2.598151
        view   sqft_above sqft_basement     yr_built
    1.050604     2.744065     1.899141     1.795086
```

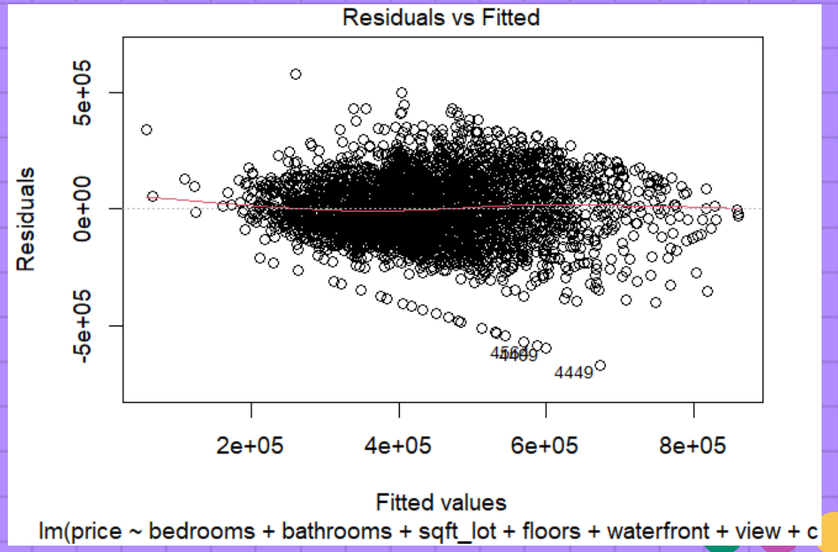
```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
    view + sqft_above + sqft_basement + yr_built, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-720171  -96914   -3565   93465  560763

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.996e+06  2.049e+05  19.499  < 2e-16 ***
bedrooms      -3.719e+04  4.032e+03  -9.225  < 2e-16 ***
bathrooms      4.383e+04  6.350e+03   6.902 6.09e-12 ***
sqft_lot      -6.932e+00  9.577e-01  -7.238 5.61e-13 ***
floors         2.798e+04  7.162e+03   3.907 9.54e-05 ***
view           2.633e+04  4.972e+03   5.296 1.26e-07 ***
sqft_above     1.813e+02  6.766e+00  26.803  < 2e-16 ***
sqft_basement  1.628e+02  9.010e+00  18.074  < 2e-16 ***
yr_built      -1.952e+03  1.070e+02 -18.237  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140200 on 3355 degrees of freedom
Multiple R-squared:  0.3842,    Adjusted R-squared:  0.3827
F-statistic: 261.6 on 8 and 3355 DF,  p-value: < 2.2e-16
```



MODEL KELIMA

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_lot + floors +
    waterfront + view + condition + sqft_above + sqft_basement +
    yr_built + yr_renovated + bedrooms:sqft_lot + bedrooms:floors +
    bedrooms:view + bedrooms:sqft_above + bedrooms:yr_built +
    bathrooms:floors + bathrooms:view + sqft_lot:floors + sqft_lot:condition +
    sqft_lot:sqft_above + sqft_lot:sqft_basement + sqft_lot:yr_built +
    floors:condition + floors:sqft_above + floors:yr_built +
    view:yr_built + condition:yr_built + condition:yr_renovated +
    sqft_above:sqft_basement + sqft_basement:yr_built, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-673256  -91539   -5430   92448  577786

Coefficients: (4 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.841e+06  1.371e+06   2.802 0.005114 **
bedrooms                  9.033e+05  2.559e+05   3.529 0.000422 ***
bathrooms                 8.825e+04  1.776e+04   4.968 7.09e-07 ***
sqft_lot                 -3.105e+02  9.914e+01  -3.132 0.001752 **
floors                   -1.524e+06  5.317e+05  -2.867 0.004168 **
waterfront                1.761e+05  1.028e+05   1.712 0.086991 .
view                     -6.383e+05  3.654e+05  -1.747 0.080766 .
condition2                5.190e+06  5.573e+06   0.931 0.351797
condition3                1.295e+06  9.427e+05   1.374 0.169492
condition4                2.867e+06  1.003e+06   2.859 0.004275 **
condition5                2.374e+05  3.947e+05   0.602 0.547502
sqft_above                1.438e+02  3.931e+01   3.657 0.000259 ***
sqft_basement            -1.125e+03  5.987e+02  -1.879 0.060376 .
yr_built                 -1.968e+03  6.930e+02  -2.840 0.004532 **
yr_renovated              3.175e+01  1.131e+01   2.807 0.005029 **
bedrooms:sqft_lot         3.091e+00  1.480e+00   2.088 0.036886 *
bedrooms:floors           3.444e+04  1.090e+04   3.158 0.001601 **
bedrooms:view             1.598e+04  7.254e+03   2.203 0.027662 *
bedrooms:sqft_above      -2.701e+01  8.193e+00  -3.297 0.000988 ***
bedrooms:yr_built        -4.919e+02  1.336e+02  -3.683 0.000234 ***
bathrooms:floors         -3.155e+04  1.100e+04  -2.868 0.004156 **
bathrooms:view           -1.724e+04  9.467e+03  -1.821 0.068649 .
sqft_lot:floors          -1.395e+01  2.210e+00  -6.314 3.08e-10 ***
sqft_lot:condition2       -2.513e+01  6.281e+01  -0.400 0.689094
sqft_lot:condition3       -3.101e+01  6.008e+01  -0.516 0.605789
sqft_lot:condition4       -2.146e+01  6.008e+01  -0.357 0.720996
sqft_lot:condition5       -2.568e+01  5.999e+01  -0.428 0.668687
sqft_lot:sqft_above        1.126e-02  2.114e-03   5.324 1.08e-07 ***
sqft_lot:sqft_basement    -7.090e-03  2.674e-03  -2.652 0.008049 **
sqft_lot:yr_built          1.652e-01  4.230e-02   3.906 9.58e-05 ***
floors:condition2        -1.380e+05  2.033e+05  -0.679 0.497383
floors:condition3        -5.338e+04  2.500e+04  -2.135 0.032802 *
floors:condition4        -1.260e+04  2.592e+04  -0.486 0.626847
floors:condition5                NA         NA      NA       NA
floors:sqft_above         3.728e+01  1.632e+01   2.285 0.022404 *
floors:yr_built           7.885e+02  2.737e+02   2.881 0.003991 **
view:yr_built             3.304e+02  1.909e+02   1.731 0.083527 .
condition2:yr_built      -2.540e+03  2.803e+03  -0.906 0.364942
condition3:yr_built      -5.031e+02  4.561e+02  -1.103 0.270110
condition4:yr_built      -1.366e+03  4.889e+02  -2.795 0.005221 **
condition5:yr_built              NA         NA      NA       NA
condition2:yr_renovated          NA         NA      NA       NA
condition3:yr_renovated  -3.175e+01  1.190e+01  -2.668 0.007674 **
condition4:yr_renovated  -2.645e+01  1.237e+01  -2.138 0.032587 *
condition5:yr_renovated          NA         NA      NA       NA
sqft_above:sqft_basement  4.103e-02  1.818e-02   2.257 0.024068 *
sqft_basement:yr_built    6.473e-01  3.124e-01   2.072 0.038341 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 136700 on 3321 degrees of freedom
Multiple R-squared:  0.4206,    Adjusted R-squared:  0.4132
F-statistic: 57.39 on 42 and 3321 DF,  p-value: < 2.2e-16
```





THANK YOU

www.reallygreatsite.com