House Price Regression

Anggota Kelompok:

Angelyne	2306209624
Andini Chiquita Dewi	2306220564
Subhan Irsyaduddien Alhaq	2306215564
Jeremy Azriel Rafael	2306217140
Faris Ahmad Zubair Bauw	2306217960

2024

Contents

1 Pendahuluan

1.1 Latar Belakang

Perumahan merupakan salah satu kebutuhan primer manusia yang terus berkembang seiring dengan peningkatan populasi dan urbanisasi. Di era modern ini, rumah tidak hanya berfungsi sebagai tempat tinggal, tetapi juga menjadi simbol stabilitas finansial dan investasi jangka panjang. Harga sebuah rumah sangat bervariasi, dipengaruhi oleh berbagai faktor internal maupun eksternal.

Faktor-faktor seperti luas bangunan, jumlah kamar tidur, jumlah kamar mandi, tahun pembangunan, ukuran lahan, kapasitas garasi, serta kualitas lingkungan sekitar berkontribusi besar dalam menentukan nilai properti. Sebagai contoh, rumah dengan luas bangunan yang besar dan lokasi di lingkungan yang berkualitas tinggi cenderung memiliki harga yang lebih tinggi dibandingkan rumah dengan karakteristik yang lebih sederhana.

Pemahaman tentang faktor-faktor yang memengaruhi harga rumah menjadi penting, baik bagi pembeli, penjual, maupun pengembang properti. Analisis mendalam terhadap data harga rumah dapat membantu pihak-pihak terkait dalam mengambil keputusan yang lebih baik, seperti menentukan harga jual, memilih lokasi pembangunan, atau mencari properti yang sesuai dengan anggaran. Oleh karena itu, penelitian ini dilakukan untuk mengeksplorasi hubungan antara berbagai karakteristik rumah, seperti luas bangunan, jumlah kamar tidur dan kamar mandi, tahun pembangunan, ukuran lahan, kapasitas garasi, serta kualitas lingkungan sekitar, terhadap harga rumah.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah:

1. Bagaimana pengaruh luas bangunan, jumlah kamar tidur, jumlah kamar mandi, tahun pembangunan, ukuran lahan, kapasitas garasi, dan kualitas lingkungan sekitar terhadap harga rumah?
2. Apakah semua variabel prediktor yang digunakan signifikan dalam memprediksi harga rumah?

1.3 Tujuan

Penelitian ini bertujuan untuk:

1. Menganalisis hubungan antara luas bangunan, jumlah kamar tidur, jumlah kamar

mandi, tahun pembangunan, ukuran lahan, kapasitas garasi, serta kualitas lingkungan sekitar terhadap harga rumah.

2. Mengidentifikasi variabel-variabel prediktor yang signifikan dalam mempengaruhi harga rumah.

1.4 Informasi Data

Data yang akan diproses memiliki 1000 baris dengan 7 variabel prediktor (luas bangungan, jumlah kamar tidur, jumlah kamar mandi, tahun pembangunan, ukuran lahan, kapasitas garasi, dan kualitas sekitar rumah) dan 1 variabel respon (harga rumah). Seluruh variabel (kolom) pada dataset tidak ada outlier dan data osong. Penjelasan untuk setiap variabel yaitu,:

- **Square_Footage**: Ukuran rumah dalam satuan kaki persegi; datanya berkisar antara 503 – 4999. (tipe data: rasio)
- **Num_Bedrooms**: Jumlah kamar tidur dalam rumah; datanya berkisar antara 1 – 5. (tipe data: rasio)
- **Num_Bathrooms**: Jumlah kamar mandi dalam rumah; datanya berkisar antara 1 – 3. (tipe data: rasio)
- **Year_Built**: Tahun pembangunan rumah; datanya berkisar antara 1950 – 2022. (tipe data: interval)
- **Lot_Size**: Ukuran lahan tempat rumah dibangun, diukur dalam satuan hektar; datanya berkisar antara 0.51 – 4.99. (tipe data: rasio)
- **Garage_Size**: Kapasitas garasi, yaitu jumlah mobil yang dapat dimuat dalam garasi; datanya berkisar antara 0 – 2. (tipe data: rasio)
- **Neighborhood_Quality**: Penilaian kualitas lingkungan rumah pada skala 1 – 10, di mana 10 menunjukkan lingkungan berkualitas tinggi. (tipe data: ordinal)
- **House_Price (Variabel Target)**: Harga rumah dalam dollar, yang merupakan variabel dependen yang ingin diprediksi; datanya berkisar antara 112,000 – 1,110,000. (tipe data: rasio)

Sumber data: `https://www.kaggle.com/datasets/prokshitha/home-value-insights/data`

2 Pre-processing (jika ada) dan Analisis Deskriptif

2.1 Meng-import Data

Hal pertama yang akan dilakukan adalah mengunduh dataset dari sumber data (kaggle) lalu dilanjut dengan meng-import data csv ke dalam software, pada kali ini akan digunakan software R-Studio.

	Square_Footage	Num_Bedrooms	Num_Bathrooms	Year_Built	Lot_Size	Garage_Size	Neighborhood_Quality	House_Price
1	1360	2	1	1981	0.5996366	0	5	262382.9
2	4272	3	3	2016	4.7530138	1	6	985260.9
3	3592	1	2	2016	3.6348227	0	9	777977.4
4	966	1	2	1977	2.7306669	1	8	229698.9
5	4926	2	1	1993	4.6990726	0	8	1041740.9
6	3944	5	3	1990	2.4759300	2	8	879797.0
7	3671	1	2	2012	4.9119601	0	1	814427.9

Figure 1: Tampilan tabular dataset House Price Regression.

2.2 Mengecek Missing Value

```
#Mengecek apakah ada missing values pada tiap variabel
which(is.na(house_price_regression$Square_Footage))
which(is.na(house_price_regression$Num_Bedrooms))
which(is.na(house_price_regression$Num_Bathrooms))
which(is.na(house_price_regression$Year_Built))
which(is.na(house_price_regression$Lot_Size))
which(is.na(house_price_regression$Garage_Size))
which(is.na(house_price_regression$Neighborhood_Quality))
which(is.na(house_price_regression$House_Price))
```

```
> #Mengecek apakah ada missing values pada tiap variabel
> which(is.na(house_price_regression$Square_Footage))
integer(0)
> which(is.na(house_price_regression$Num_Bedrooms))
integer(0)
> which(is.na(house_price_regression$Num_Bathrooms))
integer(0)
> which(is.na(house_price_regression$Year_Built))
integer(0)
> which(is.na(house_price_regression$Lot_Size))
integer(0)
> which(is.na(house_price_regression$Garage_Size))
integer(0)
> which(is.na(house_price_regression$Neighborhood_Quality))
integer(0)
> which(is.na(house_price_regression$House_Price))
integer(0)
```

Figure 2: Pengecekan missing values pada data.

Pada Gambar 2.2, didapat "integer(0)" untuk masing-masing data antara variabel prediktor dengan variabel respon yang menunjukkan bahwa tidak terdapat missing value dalam semua data, sehingga peneliti dapat melanjutkan langkah pre-processing berikutnya.

2.3 Analisis Deskriptif

Pada tahap analisis deskriptif, akan dilakukan visualisasi data yang telah melalui tahap pre-processing untuk mempermudah pengajuan hipotesis dan pemodelan pada tahapan selanjutnya. Visualisasi yang digunakan oleh peneliti adalah scatter plot dan heat map.

```
> summary(house_price_regression)
 Square_Footage  Num_Bedrooms   Num_Bathrooms     Year_Built      Lot_Size
Min.   : 503   Min.   :1.00   Min.   :1.000   Min.   :1950   Min.   :0.5061
1st Qu.:1750   1st Qu.:2.00   1st Qu.:1.000   1st Qu.:1969   1st Qu.:1.6659
Median :2862   Median :3.00   Median :2.000   Median :1986   Median :2.8097
Mean   :2815   Mean   :2.99   Mean   :1.973   Mean   :1987   Mean   :2.7781
3rd Qu.:3850   3rd Qu.:4.00   3rd Qu.:3.000   3rd Qu.:2004   3rd Qu.:3.9233
Max.   :4999   Max.   :5.00   Max.   :3.000   Max.   :2022   Max.   :4.9893
 Garage_Size    Neighborhood_Quality  House_Price
Min.   :0.000   Min.   : 1.000       Min.   : 111627
1st Qu.:0.000   1st Qu.: 3.000       1st Qu.: 401648
Median :1.000   Median : 6.000       Median : 628267
Mean   :1.022   Mean   : 5.615       Mean   : 618861
3rd Qu.:2.000   3rd Qu.: 8.000       3rd Qu.: 827141
Max.   :2.000   Max.   :10.000       Max.   :1108237
```

Figure 3: Informasi singkat setiap variabel.

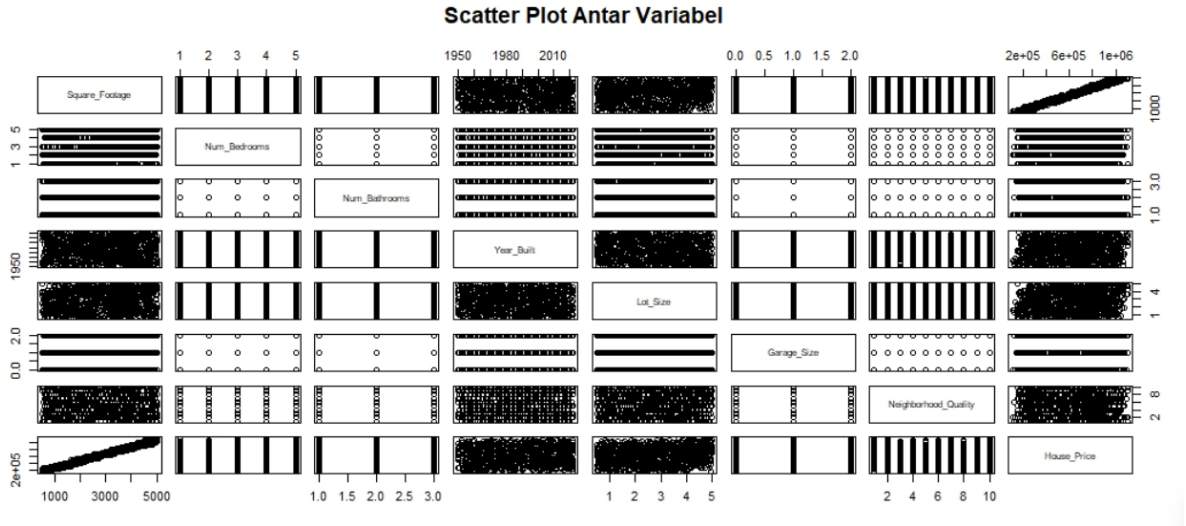

Figure 4: Pair plot antar variabel.

Gambar 2.3.2 adalah pair plot yang mem-plot seluruh variabel prediktor terhadap variabel respon dalam bentuk matriks scatter plot. Pair plot adalah visualisasi yang berguna dalam analisis data multivariat untuk memahami hubungan antara pasangan variabel dalam suatu set observasi. Pair plot tersebut digunakan sebagai alat eksplorasi awal data sebelum melakukan analisis statistik yang lebih mendalam atau membangun model prediktif.

Berikut beberapa kegunaan pair plot yang berguna untuk membangun model prediktif bagi observasi tersebut:

1. Analisis korelasi: berguna untuk memvisualisasikan sejauh mana variabel-variabel dalam dataset berkorelasi satu sama lain. Peneliti dapat melihat sejauh mana hubungan antara variabel-variabel positif (mengarah ke atas dan ke kanan) atau negatif (mengarah ke bawah dan ke kiri).

2. Analisis regresi: digunakan untuk memeriksa hubungan antara variabel independen dan variabel dependen agar dapat melihat apakah terdapat hubungan linier atau non-linier antara variabel-variabel tersebut.

3. Distribusi data: berguna pada saat ingin melihat gambaran distribusi masing-masing variabel pada diagonal untuk menilai apakah data mengikuti distribusi tertentu, seperti distribusi normal.

```
> cor(house_price_regression)
                    Square_Footage Num_Bedrooms Num_Bathrooms   Year_Built
Square_Footage          1.000000000 -0.043563949  -0.031583568 -0.022392382
Num_Bedrooms           -0.043563949  1.000000000   0.022848019 -0.015819695
Num_Bathrooms          -0.031583568  0.022848019   1.000000000 -0.021062863
Year_Built             -0.022392382 -0.015819695  -0.021062863  1.000000000
Lot_Size                0.089478617 -0.009355012   0.034923122 -0.061050327
Garage_Size             0.030592952  0.113760956   0.024845812 -0.025484503
Neighborhood_Quality   -0.008357227 -0.049024381   0.017584698 -0.009548995
House_Price             0.991261451  0.014633276  -0.001862098  0.051967362
                        Lot_Size  Garage_Size Neighborhood_Quality  House_Price
Square_Footage       0.089478617  0.030592952          -0.008357227  0.991261451
Num_Bedrooms        -0.009355012  0.113760956          -0.049024381  0.014633276
Num_Bathrooms        0.034923122  0.024845812           0.017584698 -0.001862098
Year_Built          -0.061050327 -0.025484503          -0.009548995  0.051967362
Lot_Size             1.000000000  0.002436104           0.037630329  0.160411691
Garage_Size          0.002436104  1.000000000          -0.011286860  0.052133259
Neighborhood_Quality 0.037630329 -0.011286860           1.000000000 -0.007770031
House_Price          0.160411691  0.052133259          -0.007770031  1.000000000
```

Figure 5: Korelasi antar variabel.

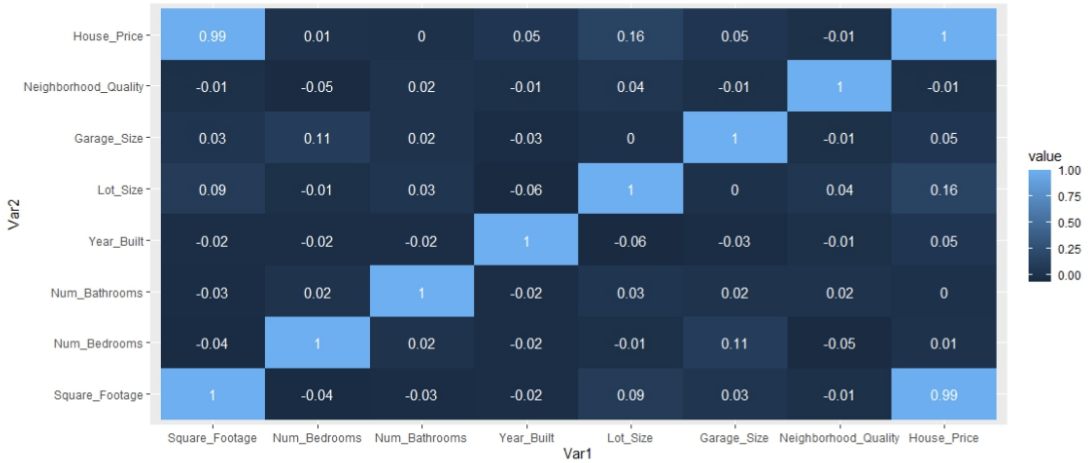

Figure 6: Heatmap dari korelasi antar variabel.

Square Footage (ukuran rumah) Ukuran rumah yang lebih besar biasanya memiliki harga lebih tinggi karena memberikan ruang yang lebih luas bagi penghuninya. Hal ini meningkatkan kenyamanan dan fungsi, yang menjadi nilai tambah signifikan bagi calon pembeli. Pembeli cenderung bersedia membayar lebih untuk rumah dengan luas lebih

besar karena lebih fleksibel untuk ditata dan dihuni. Berdasarkan heatmap, dapat dilihat bahwa square footage memiliki korelasi yang sangat tinggi yaitu 0,99 (sangat kuat).

Num Bedroom (jumlah kamar tidur) Semakin banyak kamar tidur, semakin tinggi potensi harga rumah karena rumah dapat menampung lebih banyak penghuni atau digunakan untuk tujuan keluarga besar. Rumah dengan lebih banyak kamar tidur sering dicari oleh keluarga besar atau pembeli yang memerlukan ruang tambahan untuk tamu, kantor, atau keperluan lainnya. Berdasarkan heatmap, dapat dilihat bahwa num bedrooms memiliki korelasi yang sangat lemah yaitu 0,01 (sangat lemah).

Num Bathrooms (jumlah kamar mandi) Rumah dengan lebih banyak kamar mandi cenderung lebih mahal karena memberikan kenyamanan dan efisiensi yang lebih baik bagi penghuni. Pembeli lebih tertarik pada rumah dengan jumlah kamar mandi yang memadai, terutama untuk keluarga besar yang memerlukan akses sanitasi lebih efisien dan tidak terbatas. Berdasarkan heatmap, dapat dilihat bahwa num bathrooms tidak memiliki korelasi sama sekali terhadap harga rumah.

Year Built (tahun pembangunan) Tahun pembangunan rumah menunjukkan usia bangunan. Rumah yang dibangun lebih baru cenderung memiliki desain modern dan struktur yang lebih baik. Rumah yang lebih tua umumnya memiliki harga lebih rendah karena kemungkinan adanya penurunan kualitas struktur dan fasilitas akibat usia bangunan. Sementara itu, rumah yang lebih baru cenderung lebih mahal karena masih dalam kondisi baik dan sering kali memiliki desain serta teknologi modern. Pembeli biasanya mencari rumah yang lebih baru untuk meminimalkan biaya renovasi dan perawatan. Berdasarkan heatmap, dapat dilihat bahwa year built memiliki korelasi yang sangat lemah yaitu 0,05 (sangat lemah).

Lot Size (ukuran lahan) Lahan yang lebih besar cenderung meningkatkan harga rumah karena memberikan ruang tambahan untuk taman, halaman, atau potensi pengembangan di masa depan. Ukuran lahan besar sering dicari oleh pembeli yang menginginkan privasi lebih atau fasilitas tambahan seperti kolam renang, taman bermain, atau perluasan bangunan. Berdasarkan heatmap, dapat dilihat bahwa lot size memiliki korelasi yang lemah yaitu 0,16 (lemah).

Garage Size (kapasitas garasi) Kapasitas garasi mengukur berapa banyak mobil yang dapat ditampung di garasi rumah. Rumah dengan garasi yang lebih besar cenderung lebih mahal karena menawarkan fasilitas tambahan untuk parkir kendaraan serta

penyimpanan barang. Adanya garasi memberikan kenyamanan, keamanan, dan fungsionalitas lebih bagi pemilik kendaraan, sehingga meningkatkan nilai jual rumah. Berdasarkan heatmap, dapat dilihat bahwa garage size memiliki korelasi yang sangat lemah yaitu 0,05 (sangat lemah).

Neighborhood Quality (kualitas lingkungan) Rumah yang berada di lingkungan dengan kualitas tinggi cenderung memiliki harga lebih mahal. Faktor seperti keamanan, akses ke fasilitas umum (sekolah, pusat perbelanjaan, rumah sakit), kebersihan, dan ketenangan lingkungan menjadi daya tarik utama. Pembeli bersedia membayar lebih untuk tinggal di lingkungan yang nyaman, aman, dan strategis. Berdasarkan heatmap, dapat dilihat bahwa neighborhood quality memiliki korelasi yang sangat lemah yaitu -0,01 (sangat lemah); perhatikan juga bahwa korelasi neighborhood quality merupakan korelasi negatif.

Berdasarkan visualisasi data yang telah dilakukan sebelumnya, dapat diajukan hipotesis bahwa variabel prediktor, yaitu "square footage", "num_bedrooms", "year_built", "lot_size", "garage size" dan "neighborhood quality" memiliki pengaruh dalam menjelaskan variabel respon, yaitu "price" atau harga rumah. Meskipun pada heatmap variabel num_bathrooms tidak memiliki korelasi secara langsung terhadap house_price tetapi interaksi antara num_bathrooms terhadap variabel lainnya dapat menambah keakuratan model dalam memprediksi house_price.

3 Pemodelan

Kami mengajukan 2 model regresi untuk dataset ini. Kami tidak melakukan transformasi variabel pada model terbaik yang kami gunakan karena semua asumsi terpenuhi pada bagian uji asumsi model.

3.1 Model yang Diajukan

A. Model 1

Model 1 menggunakan ketujuh variabel dalam ordo pertama saja yang digunakan sebagai variabel kuantitatif. Variabel "Num_Bedrooms", "Num_Bathrooms", "Garage_Size" memiliki skala berupa kategori yang dapat diinterpretasikan secara numerik, yaitu jarak atau selisih antar tiap skala (seperti 1 dengan 2 dan 2 dengan 3) memiliki makna yang sama, contohnya 3 kamar tidur lebih banyak 1 unit daripada 2 kamar tidur. Variabel "Neighborhood_Quality" kami asumsikan juga dapat diinterpretasikan secara numerik, yaitu digunakan suatu pengukuran tetap dalam penetapan skalanya sehingga jarak antar skala bermakna sama. Model 1:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

Penjelasan notasi:

- y = Harga rumah dalam dollar (House_Price)
- β_0 = Intersep model
- x_1 = Ukuran rumah dalam satuan kaki persegi (Square_Footage)
- x_2 = Jumlah kamar tidur dalam rumah (Num_Bedrooms)
- x_3 = Jumlah kamar mandi dalam rumah (Num_Bathrooms)
- x_4 = Tahun pembangunan rumah (Year_Built)
- x_5 = Ukuran lahan tempat rumah dibangun, diukur dalam satuan hektar (Lot_Size)
- x_6 = Kapasitas garasi, yaitu jumlah mobil yang dapat dimuat dalam garasi (Garage_Size)
- x_7 = Penilaian kualitas lingkungan rumah (Neighborhood_Quality)

Asumsi:
$\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$, error model bersifat independen dan berdistribusi normal dengan mean 0 dan variansi konstan untuk semua observasi.

```
> summary(first_order_model)

Call:
lm(formula = House_Price ~ Square_Footage + Num_Bedrooms + Num_Bathrooms +
    Year_Built + Lot_Size + Garage_Size + Neighborhood_Quality,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
-32029  -6542     33   6737  32145

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.007e+06  3.006e+04 -66.775   <2e-16 ***
Square_Footage       1.998e+02  2.485e-01 803.789   <2e-16 ***
Num_Bedrooms         1.017e+04  2.192e+02  46.409   <2e-16 ***
Num_Bathrooms        8.245e+03  3.787e+02  21.771   <2e-16 ***
Year_Built           9.915e+02  1.507e+01  65.812   <2e-16 ***
Lot_Size             1.492e+04  2.406e+02  62.013   <2e-16 ***
Garage_Size          5.158e+03  3.833e+02  13.455   <2e-16 ***
Neighborhood_Quality 8.062e+01  1.076e+02   0.749    0.454
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9799 on 992 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9985
F-statistic: 9.543e+04 on 7 and 992 DF,  p-value: < 2.2e-16
```

Figure 7: Ringkasan Model 1.

B. Model 2

Model 2 didapatkan dengan melakukan stepwise regression pada model 1. Hasilnya adalah model yang lebih sederhana karena variabel "Neighborhood_Quality" tidak dipertahankan/tidak signifikan (p-value = 0.454).

Model 2:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Penjelasan notasi:

- y = Harga rumah dalam dollar (House_Price)
- β_0 = Intersep model
- x_1 = Ukuran rumah dalam satuan kaki persegi (Square_Footage)
- x_2 = Jumlah kamar tidur dalam rumah (Num_Bedrooms)
- x_3 = Jumlah kamar mandi dalam rumah (Num_Bathrooms)
- x_4 = Tahun pembangunan rumah (Year_Built)

- x_5 = Ukuran lahan tempat rumah dibangun, diukur dalam satuan hektar (Lot_Size)
- x_6 = Kapasitas garasi, yaitu jumlah mobil yang dapat dimuat dalam garasi (Garage_Size)

Asumsi:
$\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$, error model bersifat independen dan berdistribusi normal dengan mean 0 dan variansi konstan untuk semua observasi.

```
> summary(stepwise_model)

Call:
lm(formula = House_Price ~ Square_Footage + Num_Bedrooms + Num_Bathrooms +
    Year_Built + Lot_Size + Garage_Size, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-31910  -6456    -21   6633  32110

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.006e+06  3.004e+04  -66.80   <2e-16 ***
Square_Footage  1.998e+02  2.484e-01  804.03   <2e-16 ***
Num_Bedrooms    1.016e+04  2.188e+02   46.44   <2e-16 ***
Num_Bathrooms   8.249e+03  3.786e+02   21.79   <2e-16 ***
Year_Built      9.914e+02  1.506e+01   65.82   <2e-16 ***
Lot_Size        1.493e+04  2.404e+02   62.10   <2e-16 ***
Garage_Size     5.156e+03  3.833e+02   13.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9797 on 993 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9985
F-statistic: 1.114e+05 on 6 and 993 DF,  p-value: < 2.2e-16
```

Figure 8: Ringkasan Model 1.

3.2 Perbandingan Model

Kami membandingkan kedua model di atas untuk mengetahui apakah menambahkan variabel "Neighborhood_Quality" akan meningkatkan performa model dalam memprediksi variabel target. Uji nested models dilakukan dengan,

A. Hipotesis
$H_0 : \beta_7 = 0$
$H_1 : \beta_7 \neq 0$

B. Tingkat signifikansi $\alpha = 0.05$

C. Statistik uji F tabel anova model

D. Keputusan untuk menolak H_0
jika p-value lebih kecil daripada tingkat signifikansi (α)

```
> anova(stepwise_model,first_order_model)
Analysis of Variance Table

Model 1: House_Price ~ Square_Footage + Num_Bedrooms + Num_Bathrooms +
    Year_Built + Lot_Size + Garage_Size
Model 2: House_Price ~ Square_Footage + Num_Bedrooms + Num_Bathrooms +
    Year_Built + Lot_Size + Garage_Size + Neighborhood_Quality
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    993 9.5303e+10
2    992 9.5249e+10  1  53880888 0.5612  0.454
```

Figure 9: Hasil Anova.

Didapatkan p-value = 0.454.

Karena p-value lebih besar daripada tingkat signifikansi (0.454 ¿ 0.05), maka H_0 gagal ditolak. Kesimpulannya menambahkan variabel "Neighborhood_Quality" ke dalam model tidak meningkatkan performa model dalam memprediksi variabel target "House_Price".

Berdasarkan uji tersebut, model 2 lebih baik digunakan karena lebih sederhana dan memiliki performa yang sangat baik. Kemudian diputuskan variabel "Neighborhood_Quality" tidak memiliki peran signifikan dalam memprediksi variabel target.

3.3 Uji Asumsi Model

Model yang kami gunakan (model 2) perlu memenuhi beberapa asumsi model regresi yang dapat diujikan yaitu:

1. **Asumsi linearitas**
   Hubungan antara variabel independen (X) dengan variabel dependen (Y) harus linear. Pada plot "Residuals vs Fitted Values" terlihat residual tersebar secara acak di sekitar garis horizontal (y = 0). Sehingga model 2 memenuhi asumsi linearitas.

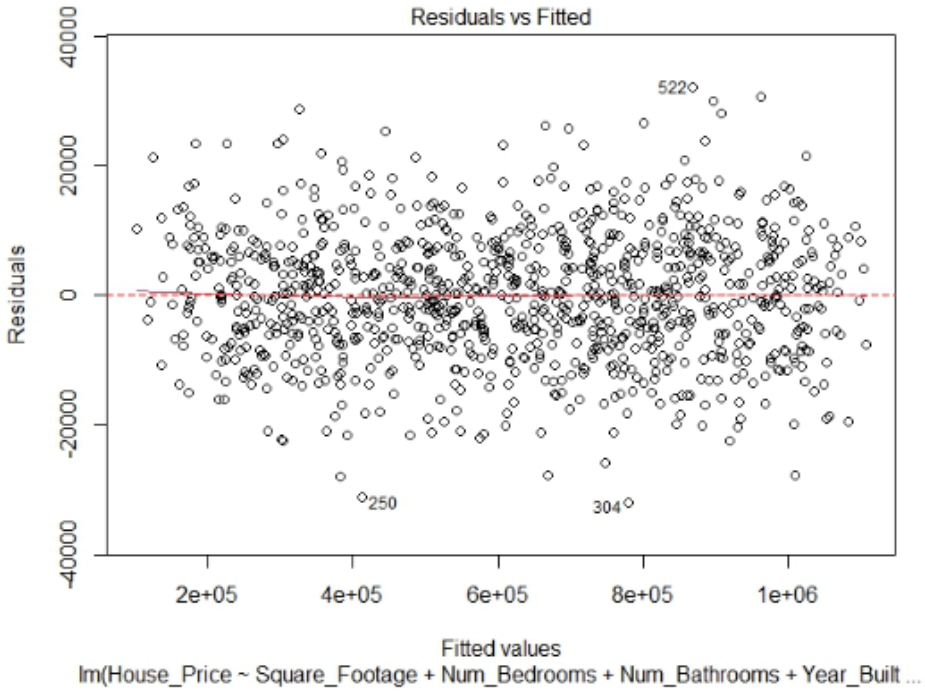

Figure 10: Plot Residuals vs Fitted Values.

2. **Asumsi independensi error**

   Tidak ada autokorelasi. Asumsi ini diuji menggunakan Durbin–Watson Test. Jika p-value lebih besar dari α (0.05), maka diputuskan tidak ada autokorelasi. Didapatkan p-value = 0.282, sehingga model 2 memenuhi asumsi independensi error.

3. **Asumsi homoskedastisitas**

   Variansi error konstan untuk semua nilai (X). Pada plot "Scale–Location" terlihat titik-titik pada plot tersebar secara acak dan tidak membentuk suatu pola tertentu. Asumsi homoskedastisitas juga diuji dengan Breusch–Pagan Test dimana p-value lebih besar dari α (0.05) menunjukkan tidak adanya heteroskedastisitas. Berdasarkan plot "Scale–Location" dan Breusch–Pagan Test (p-value = 0.458) kami simpulkan model 2 memenuhi asumsi homoskedastisitas.

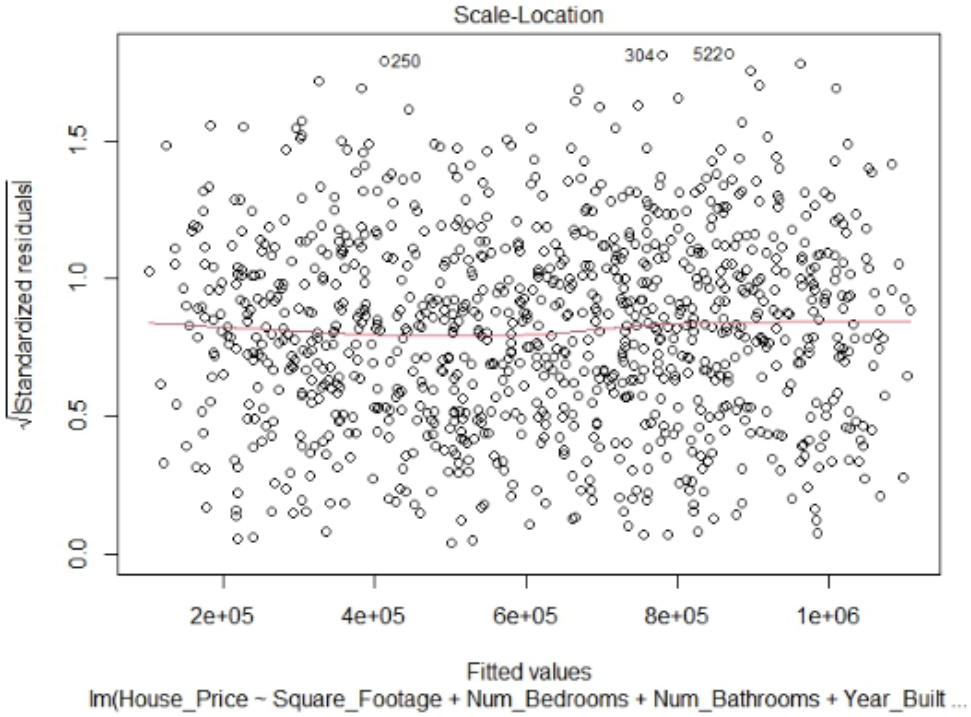

Figure 11: Plot Scale-Location.

4. **Asumsi normalitas**

   Error berdistribusi normal. Pada plot "Q–Q Residuals" terlihat titik-titik mengikuti atau berada di sekitar garis lurus. Hal ini menunjukkan model 2 memenuhi asumsi normalitas. Asumsi ini juga dapat diuji dengan Shapiro–Wilk Test dimana p-value yang besar menunjukkan asumsi normalitas terpenuhi. Didapatkan p-value = 0.6255 sehingga model 2 memenuhi asumsi normalitas.

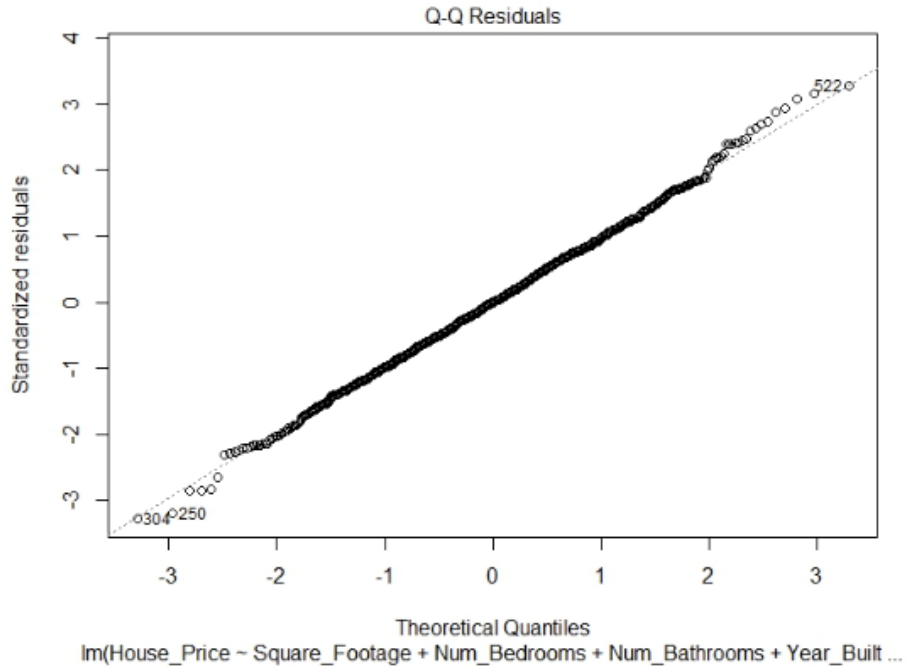

Figure 12: Plot Q–Q Residuals.

5. **Asumsi tidak ada multikolinearitas**
   Variabel independen (X) tidak saling berkorelasi secara sempurna. Pada heatmap "Korelasi antar variabel", masing-masing variabel independen tidak memiliki koefisien korelasi yang tinggi dengan variabel lain. Hal ini menunjukkan tidak ada multikolinearitas yang terjadi. Asumsi ini dapat diuji dengan melihat nilai VIF (Variance Inflation Factor) dimana nilai VIF yang lebih kecil dari 10 menunjukkan tidak ada multikolinearitas pada masing-masing variabel independen. Semua variabel menunjukkan nilai VIF yang kecil (sekitar 1) sehingga diputuskan model 2 memenuhi asumsi tidak adanya multikolinearitas.

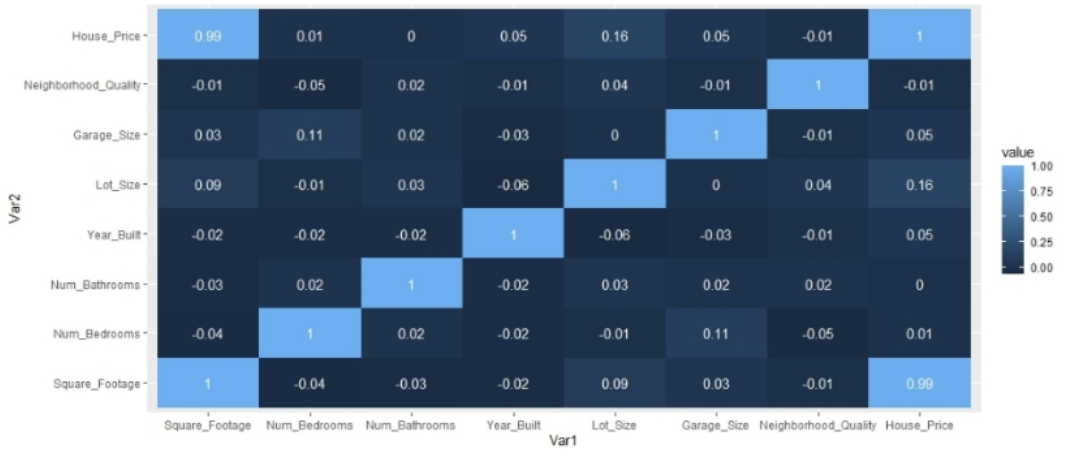

Figure 13: Heatmap dari korelasi antar variabel..

```
> # 1. Linearitas (Residuals vs Fitted Values Plot)
> plot(stepwise_model, which = 1)  # Residuals vs Fitted
> abline(h = 0, col = "red", lty = 2)  # Garis horizontal pada 0
>
> # 2. Independensi Residual (Durbin-Watson Test)
> dw_test <- durbinWatsonTest(stepwise_model)
> print(dw_test)  # p-value > 0.05 (no autocorrelation)
 lag Autocorrelation D-W Statistic p-value
   1     -0.03513998      2.067261   0.282
 Alternative hypothesis: rho != 0
>
> # 3. Homoskedastisitas (Scale-Location Plot)
> plot(stepwise_model, which = 3)
> bp_test <- bptest(stepwise_model) # Breusch-Pagan test
> print(bp_test)  # p-value > 0.05 (no heteroscedasticity)

        studentized Breusch-Pagan test

data:  stepwise_model
BP = 5.6971, df = 6, p-value = 0.458

>
> # 4. Normalitas Residual (Q-Q Plot and Shapiro-Wilk Test)
> plot(stepwise_model, which = 2)
> shapiro_test <- shapiro.test(residuals(stepwise_model))
> print(shapiro_test)  #p-value > 0.05 (normal residuals)

        Shapiro-Wilk normality test

data:  residuals(stepwise_model)
W = 0.99861, p-value = 0.6255

>
> # 5. Multikolinearitas untuk Variance Inflation Factor (VIF)
> vif_values <- vif(stepwise_model)
> print(vif_values)  # VIF values < 10 (no multicolinearity)
Square_Footage   Num_Bedrooms  Num_Bathrooms      Year_Built       Lot_Size    Garage_Si
      1.012749       1.015983       1.003819       1.005214       1.013050       1.0154
```

Figure 14: Hasil Variance Inflation Factor.

4 Pengolahan Data dan Analisis Hasil

Bagian 2 mengajukan hipotesis bahwa semua variabel prediktor ("square footage", "num_bedrooms", "year_built", "lot_size", "garage size" dan "neighborhood quality") memiliki pengaruh dalam menjelaskan variabel respon "price". Pada bagian ini, hipotesis tersebut akan diujikan menggunakan statistik uji t. Harapan dari pengujian hipotesis ini adalah menemukan adanya indikasi variabel prediktor yang berpengaruh secara signifikan terhadap variabel respon, agar dapat dilakukan analisis lebih lanjut pada data.

Menggunakan software R, dan data hasil preprocessing pada bagian 2, diperoleh hasil pengujian hipotesis sebagai berikut (kode dilampirkan di bagian Lampiran).

```
Call:
lm(formula = House_Price ~ Square_Footage + Num_Bedrooms + Num_Bathrooms +
    Year_Built + Lot_Size + Garage_Size + Neighborhood_Quality,
    data = house_price_regression)

Residuals:
   Min     1Q Median     3Q    Max
-32029  -6542     33   6737  32145

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.007e+06  3.006e+04 -66.775   <2e-16 ***
Square_Footage       1.998e+02  2.485e-01 803.789   <2e-16 ***
Num_Bedrooms         1.017e+04  2.192e+02  46.409   <2e-16 ***
Num_Bathrooms        8.245e+03  3.787e+02  21.771   <2e-16 ***
Year_Built           9.915e+02  1.507e+01  65.812   <2e-16 ***
Lot_Size             1.492e+04  2.406e+02  62.013   <2e-16 ***
Garage_Size          5.158e+03  3.833e+02  13.455   <2e-16 ***
Neighborhood_Quality 8.062e+01  1.076e+02   0.749    0.454
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9799 on 992 degrees of freedom
Multiple R-squared:  0.9985,    Adjusted R-squared:  0.9985
F-statistic: 9.543e+04 on 7 and 992 DF,  p-value: < 2.2e-16
```

Figure 15: Output uji t pada model regresi.

Pada poin sebelumnya semua variabel diujikan. Berdasarkan nilai p-value yang diperoleh, semua variabel terindikasi berpengaruh terhadap variabel respon, kecuali pada variabel "Neighborhood_Quality". Karena hipotesis yang diajukan pada bagian 2 menyatakan bahwa semua variabel prediktor berpengaruh, maka, dapat disimpulkan bahwa terdapat bukti yang cukup untuk menolak hipotesis tersebut.

Setelah melakukan stepwise regression, diperoleh model yang lebih sederhana, yaitu model 2. Untuk membuktikan bahwa model tersebut adalah model yang terbaik dilakukan model validation dengan cross-validation.

```
> # Output hasil cross-validation
> print(cv_model)
Linear Regression

1000 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 900, 900, 900, 900, 900, 900, ...
Resampling results:

  RMSE      Rsquared  MAE
  9787.108  0.998514  7795.028

Tuning parameter 'intercept' was held constant at a value of TRUE
> # Evaluasi performa
> cat("RMSE (Cross-validation):", cv_model$results$RMSE, "\n")
RMSE (Cross-validation): 9787.108
> cat("R-squared (Cross-validation):", cv_model$results$Rsquared, "\n")
R-squared (Cross-validation): 0.998514
```

Figure 16: Hasil cross-validation.

Dari output tersebut terlihat bahwa model ini sangat baik dengan $R^2 = 0.998514$, yang menunjukkan model dapat menjelaskan 99.85% variabilitas dalam data. RMSE dan MAE relatif kecil jika dibandingkan dengan skala data pada House_Price. Telah dilakukan uji asumsi juga pada bagian Pemodelan, sehingga dapat disimpulkan bahwa model ini adalah model terbaik dan tidak overfitting.

4.1 Uji Asumsi (Ringkas)

- **Asumsi linearitas**
  Hubungan antara variabel independen (X) dengan variabel dependen (Y) harus linear. Pada plot "Residuals vs Fitted Values" terlihat residual tersebar secara acak di sekitar garis horizontal (y = 0). Sehingga model 2 memenuhi asumsi linearitas.

- **Asumsi independensi error**
  Tidak ada autokorelasi. Asumsi ini diuji menggunakan Durbin–Watson Test. Jika p-value lebih besar dari α (0.05), maka diputuskan tidak ada autokorelasi. Didapatkan p-value = 0.282, sehingga model 2 memenuhi asumsi independensi error.

- **Asumsi homoskedastisitas**
  Variansi error konstan untuk semua nilai (X). Pada plot "Scale–Location" terlihat titik-titik pada plot tersebar secara acak dan tidak membentuk suatu pola tertentu. Asumsi homoskedastisitas juga diuji dengan Breusch–Pagan Test dimana p-value lebih besar dari α (0.05) menunjukkan tidak adanya heteroskedastisitas.

Berdasarkan plot "Scale–Location" dan Breusch–Pagan Test (p-value = 0.458) kami simpulkan model 2 memenuhi asumsi homoskedastisitas.

- **Asumsi normalitas**
  Error berdistribusi normal. Pada plot "Q–Q Residuals" terlihat titik-titik mengikuti atau berada di sekitar garis lurus. Hal ini menunjukkan model 2 memenuhi asumsi normalitas. Asumsi ini juga dapat diuji dengan Shapiro–Wilk Test dimana p-value yang besar menunjukkan asumsi normalitas terpenuhi. Didapatkan p-value = 0.6255 sehingga model 2 memenuhi asumsi normalitas.

- **Asumsi tidak ada multikolinearitas**
  Variabel independen (X) tidak saling berkorelasi secara sempurna. Pada heatmap "Korelasi antar variabel", masing-masing variabel independen tidak memiliki koefisien korelasi yang tinggi dengan variabel lain. Semua variabel menunjukkan nilai VIF yang kecil (sekitar 1) sehingga diputuskan model 2 memenuhi asumsi tidak adanya multikolinearitas.

4.2 Insight

Regresi memberikan berbagai insight yang berguna tergantung pada konteks data dan tujuan analisis. Berikut adalah beberapa informasi utama yang dapat diperoleh dari hasil regresi:

1. **R-squared (R2)**
   Nilai R-squared sebesar 0.9985 itu menunjukan 99.85% variabilitas ini dapat dijelaskan oleh model.

2. **P-value**
   Semua variabel independen memiliki nilai p-value di bawah 0.05 ini menunjukan bahwa setiap variabel independen berguna.

5 Penutup

Setelah melalui berbagai proses seperti analisis residual dan model validation, model terbaik yang diperoleh adalah sebagai berikut:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

Penjelasan notasi:

- y = Harga rumah dalam dollar (House_Price)
- β_0 = Intersep model
- x_1 = Ukuran rumah dalam satuan kaki persegi (Square_Footage)
- x_2 = Jumlah kamar tidur dalam rumah (Num_Bedrooms)
- x_3 = Jumlah kamar mandi dalam rumah (Num_Bathrooms)
- x_4 = Tahun pembangunan rumah (Year_Built)
- x_5 = Ukuran lahan tempat rumah dibangun, diukur dalam satuan hektar (Lot_Size)
- x_6 = Kapasitas garasi, yaitu jumlah mobil yang dapat dimuat dalam garasi (Garage_Size)

Asumsi:
$\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$, error model bersifat independen dan berdistribusi normal dengan mean 0 dan variansi konstan untuk semua observasi.

Hal ini disebabkan model tersebut memiliki nilai adjusted R^2 yang sangat baik, yaitu sebesar 99.85% yang berarti bahwa sekitar 99.85% variasi dari nilai variabel respons dapat dijelaskan oleh variabel prediktor. Dengan menggunakan uji F, didapatkan p-value model tersebut sebesar 2.2×10^{-16} yang lebih kecil dari nilai signifikansi sebesar 0.05. Hal ini berarti bahwa, model tersebut statistically useful. Lalu, dengan menggunakan uji t, didapat hasil p-value model yang tertera lebih kecil dari α untuk semua variabel prediktor, sehingga semua variabel prediktor memiliki pengaruh yang signifikan terhadap model yang diuji.

Model ini juga memenuhi asumsi-asumsi dari residual, yaitu asumsi bahwa ekspektasi residual bernilai 0, bersifat homoskedastisitas, berdistribusi normal, serta independen dengan residual lainnya melalui serangkaian tes pada bagian pengolahan data dan analisis hasil. Lalu, model ini juga telah tervalidasi dengan menggunakan teknik cross-validation, dimana didapat bahwa nilai R^2 melebihi model. Hal ini membuat model tersebut robust dan tervalidasi fit dengan sampel data lainnya. Oleh karena itu, dapat disimpulkan bahwa model ini adalah model terbaik dari dua model yang diajukan pada bagian pemodelan.

Dari seluruh tes di atas, terbukti bahwa hanya variabel "square footage", "num_bedrooms", "year_built", "lot_size", "garage size" memiliki pengaruh terhadap variabel respons, yaitu "price" atau harga rumah. Didapat bahwa semua variabel prediktor yang didefinisikan signifikan untuk mengestimasi harga rumah kecuali variabel prediktor "neighborhood quality". Akan tetapi, tidak dapat ditetapkan hubungan sebab akibat antar variabel respons dengan prediktor karena dataset yang digunakan merupakan data hasil observasi. Model terbaik yang didapat peneliti adalah:

$$\begin{aligned} E(y) = &- 2.006 \times 10^6 + 1.998 \times 10^2 x_1 + 1.016 \times 10^4 x_2 + 8.249 \times 10^3 x_3 + 9.914 \times 10^2 x_4 \\ &+ 1.493 \times 10^4 x_5 + 5.156 \times 10^3 x_6 \end{aligned}$$

Akan tetapi, tidak menutup kemungkinan adanya model lain yang lebih baik dibandingkan model ini karena limitasi penugasan ini yang menitikberatkan pada pengaplikasian regresi linear berganda pada dataset yang dimiliki.

6 Lampiran

Link data, kode, dan video:
`https://drive.google.com/drive/folders/1fFQmyaNS8QouGiaBSDMTD-oQLqdEW4e2?usp=drive_link`

Catatan

File gambar (mis. `gambar_2_1.png`, `output_uji_t.png`, `cross_validation.png`, dll.) bersifat opsional. Dokumen tetap dapat *compile* karena menggunakan `\safeincludegraphics`.