PREDIKSI HARGA JUAL RUMAH MENGGUNAKAN MODEL REGRESI

Makalah ini disusun untuk memenuhi Project Akhir Mata Kuliah Model Linear.

Dosen Pengampu: Madona Yunita Wijaya, M.Sc

Disusun Oleh :

Alifia Intan	11230940000016
Fairuuzzari Ramadhan	11230940000040
Aprilia Nur Afifah	11230940000060
Sekar Afifa Cettastami	11230940000062

PROGRAM STUDI MATEMATIKA FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH

JAKARTA

2025M/1446 H

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Puji syukur kami panjatkan ke hadirat Allah SWT yang telah memberikan rahmat dan karunia-Nya, sehingga kami dapat menyelesaikan laporan proyek akhir mata kuliah **Model Linear** dengan judul "**PREDIKSI HARGA JUAL RUMAH MENGGUNAKAN MODEL REGRESI**" ini dengan baik.

Laporan ini disusun sebagai bentuk penerapan konsep model linear, khususnya regresi linear, dalam menganalisis dan memprediksi harga jual rumah berdasarkan berbagai variabel karakteristik properti. Dengan menggunakan pendekatan statistik yang tepat, kami berusaha menggambarkan bagaimana model regresi dapat digunakan untuk memahami hubungan antara harga rumah dengan faktor-faktor seperti luas tanah, tipe bangunan, kondisi fisik, dan variabel lainnya.

Kami mengucapkan terima kasih yang sebesar-besarnya kepada dosen pengampu mata kuliah **Model Linear**, serta semua pihak yang telah memberikan dukungan, bimbingan, dan masukan yang berarti selama penyusunan laporan ini. Kami menyadari bahwa laporan ini masih memiliki kekurangan, oleh karena itu kritik dan saran yang membangun sangat kami harapkan demi perbaikan di masa mendatang.

Akhir kata, semoga laporan ini dapat memberikan manfaat dan menambah wawasan bagi pembaca, khususnya dalam penerapan model linear dalam bidang ekonomi dan properti.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Tangerang Selatan, 10 Juli 2025

Penulis

ABSTRAK

Cacar air merupakan salah satu penyakit infeksi menular yang umum dijumpai, khususnya pada anak-anak, dan memiliki tingkat penyebaran yang relatif tinggi dalam populasi padat. Dalam upaya memahami dinamika penyebaran penyakit ini secara kuantitatif, pendekatan berbasis model stokastik menjadi sangat relevan, mengingat adanya unsur ketidakpastian dalam proses penularan. Penelitian ini memanfaatkan konsep **Rantai Markov** sebagai alat analisis untuk memodelkan transisi status kesehatan individu dalam suatu populasi, dari kondisi rentan (susceptible) menjadi terinfeksi (infected), dan kemudian sembuh (recovered). Dengan menyusun matriks transisi berdasarkan probabilitas berpindah antarstatus pada tiap langkah waktu, dilakukan analisis terhadap perilaku sistem secara jangka panjang melalui perhitungan distribusi stasioner dan ekspektasi waktu transisi.

Hasil yang diperoleh menunjukkan bahwa model Rantai Markov mampu menggambarkan dinamika penyebaran cacar air secara realistis dalam kerangka probabilistik, serta memberikan gambaran prediktif mengenai proporsi individu dalam setiap status pada keseimbangan. Selain itu, model ini juga memungkinkan pengambilan keputusan yang lebih tepat dalam penyusunan strategi pencegahan dan penanggulangan penyebaran penyakit. Dengan demikian, penerapan teori stokastik, khususnya model Rantai Markov, terbukti menjadi pendekatan yang efektif dalam mempelajari fenomena penyebaran penyakit menular seperti cacar air secara sistematis dan terukur.

ABSTRACT

Chickenpox is one of the most common infectious diseases, particularly among children, with a relatively high transmission rate in densely populated areas. In an effort to quantitatively understand the dynamics of its spread, stochastic modeling approaches are highly relevant due to the inherent uncertainty in the transmission process. This study employs the concept of **Markov Chains** as an analytical tool to model the transition of individual health states within a population—from susceptible to infected, and eventually to recovered. By constructing a transition matrix based on the probabilities of moving between states at each time step, an analysis of the long-term behavior of the system is conducted through the calculation of stationary distributions and expected transition times.

The results indicate that the Markov Chain model is capable of realistically representing the spread of chickenpox within a probabilistic framework and provides predictive insights regarding the proportion of individuals in each state at equilibrium. Furthermore, this model supports more informed decision-making in designing strategies for the prevention and control of disease transmission. Thus, the application of stochastic theory, particularly the Markov Chain model, proves to be an effective approach in systematically and quantitatively studying the phenomenon of infectious disease spread such as chickenpox.

DAFTAR ISI

KATA PENGANTAR 2
ABSTRAK 3
DAFTAR ISI 5
BAB I 7
PENDAHULUAN 7
1.1 Latar Belakang 7
1.2 Rumusan Masalah 8
1.3 Tujuan 8
1.4 Manfaat 8
BAB II 10
LANDASAN TEORI 10
2.1 Model Linear 10
2.2 Regresi Linear 10
2.3 Asumsi Dasar Model Regresi Linear 11
2.4 Evaluasi Kebaikan Model Regresi 12
BAB III 14
METODOLOGI 14
3.1 Jenis dan Sumber Data 14
3.2 Variabel Penelitian 14
3.3 Metode Analisis Data 15
BAB IV 16
HASIL DAN PEMBAHASAN 16
4.1 Pra Pemrosesan Data 16
4.2 Eksplorasi Awal 18
4.3 Pemodelan Regresi Linear 19
4.4 Uji Asumsi Model Regresi 20
4.5 Uji Asumsi Model Regresi (Setelah Transformasi) 22
4.6 Evaluasi Model 24

4.7 Strategi Model Building: Uji Interaksi dan Polinomial 25
4.8 Perbandingan Tiga Model 26
BAB V 28
KESIMPULAN DAN SARAN 28
5.1 Kesimpulan 28
5.2 Saran 29
DAFTAR PUSTAKA 31

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perumahan merupakan salah satu kebutuhan dasar manusia yang sangat penting, selain sandang dan pangan. Dalam perkembangannya, sektor properti, khususnya perumahan, telah menjadi salah satu indikator pertumbuhan ekonomi di berbagai negara, termasuk Indonesia. Permintaan akan rumah tinggal terus meningkat seiring pertumbuhan penduduk dan urbanisasi, yang berdampak langsung terhadap naiknya harga rumah di berbagai wilayah.

Namun, penentuan harga jual rumah bukanlah proses yang sederhana. Harga sebuah rumah dipengaruhi oleh banyak faktor, baik yang bersifat fisik seperti luas tanah, jumlah kamar, dan kondisi bangunan, maupun faktor eksternal seperti lokasi, zona pemukiman (zoning), serta tahun penjualan. Kompleksitas ini menuntut adanya metode yang tepat untuk menganalisis dan memprediksi harga rumah secara objektif dan sistematis.

Dalam dunia akademik dan profesional, salah satu pendekatan yang banyak digunakan untuk menganalisis hubungan antara harga rumah dan karakteristiknya adalah **model regresi linear**. Model ini memungkinkan kita untuk melihat pengaruh satu atau lebih variabel bebas (independen) terhadap variabel terikat (dependen), yaitu harga jual rumah. Keunggulan regresi linear terletak pada kesederhanaannya, kemudahan interpretasi, dan kekuatannya dalam menangkap pola linier yang tersembunyi di balik data.

Melalui mata kuliah Model Linear, mahasiswa dibekali dengan pemahaman teoritis dan praktis mengenai bagaimana membangun model regresi, melakukan estimasi parameter, mengevaluasi kualitas model, dan melakukan interpretasi hasil. Oleh karena itu, penyusunan laporan ini bertujuan sebagai sarana penerapan langsung dari konsep-konsep tersebut dalam konteks nyata.

Dengan memanfaatkan data sekunder mengenai properti rumah, laporan ini mencoba membangun model regresi linear untuk memprediksi harga jual rumah berdasarkan karakteristik yang dimilikinya. Hasil dari pemodelan ini tidak hanya diharapkan mampu memberikan prediksi harga yang mendekati kenyataan, tetapi juga menunjukkan seberapa besar pengaruh masing-masing variabel terhadap harga tersebut.

Melalui laporan ini, diharapkan pembaca dapat memahami pentingnya pemodelan statistik dalam pengambilan keputusan di dunia nyata, serta menyadari bahwa pendekatan kuantitatif seperti model regresi dapat menjadi alat bantu yang kuat untuk menganalisis data yang kompleks dan bersifat multidimensional.

1.2 Rumusan Masalah

1. Faktor-faktor apa saja yang secara signifikan mempengaruhi variabel respon yang diteliti dan berapa besarnya?
2. Bagaimana performa model untuk prediksi?

1.3 Tujuan

1. Mengetahui variabel-variabel independen yang secara signifikan memengaruhi harga jual rumah, serta mengukur besarnya pengaruh masing-masing variabel tersebut.
2. Mengevaluasi performa model regresi yang dibangun untuk memprediksi harga jual rumah, baik secara statistik maupun interpretatif.

1.4 Manfaat

1. **Secara akademis**, penelitian ini memberikan ilustrasi nyata tentang bagaimana teori regresi linear dapat diterapkan dalam kasus dunia nyata, khususnya dalam analisis data harga rumah. Ini dapat menjadi bahan pembelajaran untuk memahami proses identifikasi variabel signifikan serta evaluasi performa model prediktif.
2. **Bagi pembuat kebijakan atau pengembang properti**, hasil dari penelitian ini dapat membantu pengembang atau analis pasar properti dalam mengenali faktor-faktor yang paling memengaruhi harga jual rumah, sehingga mereka dapat membuat keputusan strategis terkait pembangunan, penetapan harga, dan pemasaran.

3. **Bagi calon pembeli rumah atau masyarakat umum**, penelitian ini memberikan gambaran mengenai karakteristik rumah seperti apa yang secara statistik memengaruhi harga jual. Informasi ini bisa digunakan sebagai referensi dalam proses perencanaan pembelian properti.

BAB II

LANDASAN TEORI

2.1 Model Linear

Model linear adalah bentuk persamaan matematis yang digunakan untuk menggambarkan hubungan antara satu variabel dependen (respon) dengan satu atau lebih variabel independen (prediktor), di mana hubungan tersebut diasumsikan linier. Model ini banyak digunakan dalam statistika, ekonomi, dan ilmu sosial karena sifatnya yang sederhana namun cukup kuat untuk menganalisis data dan membuat prediksi.

Model linear umumnya dinyatakan dalam bentuk umum:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon$$

dimana y adalah variabel dependen (respon), $X_1, X_2, \dots, X_i$ adalah variabel independent (predictor), β_0 merupakan intercept (konstanta), $\beta_1, \beta_2, \dots, \beta_i$ adalah koefisien regresi, serta ϵ adalah residu (error).

Model linear digunakan untuk berbagai tujuan, salah satunya adalah untuk prediksi dan penjelasan hubungan antar variabel.

2.2 Regresi Linear

Regresi linear merupakan metode yang digunakan untuk menganalisis hubungan antara satu variabel respon dengan satu atau lebih variabel prediktor dengan asumsi hubungan linier. Regresi linear dibagi menjadi dua bentuk:

1. Regresi Linear Sederhana

   Hanya melibatkan satu variabel bebas.

   $$y = \beta_0 + \beta_1 X_1 + \epsilon$$

2. Regresi Linear Berganda

Melibatkan dua atau lebih variabel bebas. Model ini lebih kompleks dan mampu menjelaskan fenomena yang dipengaruhi oleh banyak faktor, seperti kasus prediksi harga rumah.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon$$

Dalam konteks penelitian ini, digunakan regresi linear berganda untuk memprediksi harga jual rumah berdasarkan banyak karakteristik properti yang tersedia dalam dataset.

2.3 Asumsi Dasar Model Regresi Linear

Agar hasil analisis regresi linear dapat diinterpretasikan secara valid dan akurat, maka terdapat beberapa asumsi dasar yang harus dipenuhi, yaitu:

1. Asumsi Linearitas

   Hubungan antara variabel independen dan dependen harus bersifat linier. Ini bisa diperiksa melalui scatter plot atau residual plot.
2. Asumsi Independensi

   Residual (galat) dari masing-masing observasi harus bersifat independen, tidak boleh saling bergantung. Hal ini penting untuk menjamin tidak adanya pola yang memengaruhi error secara sistematik.
3. Asumsi Homoskedastisitas

   Varians residual harus konstan di seluruh nilai variabel independen. Jika varians residual berubah-ubah, maka terjadi heteroskedastisitas yang bisa menyebabkan model tidak efisien.
4. Asumsi Normalitas Residual

   Residual harus mengikuti distribusi normal. Asumsi ini penting terutama jika model akan digunakan untuk membuat inferensi seperti uji signifikansi atau pembuatan interval prediksi.
5. Tidak Ada Multikolinearitas Tinggi

   Variabel independen sebaiknya tidak berkorelasi sangat tinggi satu sama lain. Jika multikolinearitas tinggi terjadi, maka interpretasi terhadap koefisien regresi menjadi tidak stabil dan membingungkan. Dapat dicek dengan VIF (Variance Inflation Factor).

Jika salah satu atau lebih dari asumsi ini dilanggar, model yang dibangun bisa memberikan hasil yang menyesatkan atau tidak akurat. Oleh karena itu, sebelum menarik kesimpulan, perlu dilakukan pengecekan dan pengujian terhadap asumsi-asumsi ini.

2.4 Evaluasi Kebaikan Model Regresi

Setelah model regresi dibangun, kita perlu mengevaluasi seberapa baik model tersebut menjelaskan data dan memberikan prediksi. Berikut adalah metrik evaluasi utama:

1. R-squared (R^2)

   R^2 adalah ukuran seberapa besar variansi dari variabel dependen yang dapat dijelaskan oleh variabel independen dalam model. Nilainya antara 0 dan 1.

   - $R^2 = 0 \Rightarrow$ model tidak menjelaskan variansi sama sekali.
   - $R^2 = 1 \Rightarrow$ model menjelaskan seluruh variansi.

2. Adjusted R-squared

   Merupakan versi R^2 yang telah disesuaikan dengan jumlah variabel prediktor. Berguna untuk membandingkan model dengan jumlah variabel berbeda. Penambahan variabel yang tidak relevan akan membuat adjusted R^2 turun, meskipun R^2 mungkin naik.

3. Mean Squared Error (MSE)

   Mengukur rata-rata dari kuadrat selisih antara nilai aktual dan nilai prediksi. Semakin kecil MSE, semakin baik model.

4. Root Mean Squared Error (RMSE)

   Merupakan akar kuadrat dari MSE. RMSE memberikan satuan yang sama dengan data awal, sehingga lebih mudah diinterpretasikan.

5. Uji Signifikansi Koefisien (Uji t)

   Digunakan untuk menguji apakah masing-masing koefisien regresi berbeda secara signifikan dari nol.

6. Uji F

   Untuk menguji apakah seluruh variabel independen secara bersama-sama signifikan dalam mempengaruhi variabel dependen.

Evaluasi ini penting agar model tidak hanya baik dari segi matematis, tetapi juga memiliki kemampuan generalisasi dan interpretasi yang kuat.

BAB III

METODOLOGI

3.1 Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder yang bersumber dari dataset harga rumah yang telah tersedia dalam format spreadsheet. Data ini mencakup berbagai karakteristik properti, seperti luas tanah (LotArea), tipe bangunan (MSSubClass), zona pemukiman (MSZoning), jumlah lantai, tipe atap, dan variabel-variabel lainnya yang diduga memiliki pengaruh terhadap harga jual rumah (*SalePrice*).

Dataset ini bersifat kuantitatif, yang berarti semua variabel direpresentasikan dalam bentuk angka atau data kategorik yang dikonversi ke bentuk numerik untuk keperluan analisis.

3.2 Variabel Penelitian

Penelitian ini terdiri atas dua jenis variabel utama:

- Variabel Dependen (Y)

  SalePrice: Harga jual rumah (dalam satuan mata uang tertentu)
- Variabel Independen (X)

  Beberapa variabel independen yang digunakan meliputi:

  1) **LotArea**: Luas tanah
  2) **OverallQual**: Kualitas keseluruhan rumah
  3) **YearBuilt**: Tahun dibangun
  4) **TotRmsAbvGrd**: Total jumlah ruangan di atas tanah
  5) **GarageArea**: Luas garasi
  6) **GrLivArea**: Luas lantai atas tanah

     dan beberapa variabel lainnya yang relevan setelah dilakukan seleksi fitur

Pemilihan variabel dilakukan berdasarkan pengetahuan domain serta hasil eksplorasi awal terhadap data (data preprocessing dan analisis korelasi).

3.3 Metode Analisis Data

Metode analisis yang digunakan dalam penelitian ini adalah regresi linear berganda, karena bertujuan untuk mengetahui pengaruh dari banyak variabel independen terhadap satu variabel dependen.

Langkah-langkah analisis yang dilakukan adalah sebagai berikut:

1. **Pra Pemrosesan Data**

- Pengecekan data kosong (missing value) dan penanganannya
- Transformasi data kategorikal menjadi data numerik (jika diperlukan)
- Deteksi dan penanganan outlier (jika ada)
- Standardisasi atau normalisasi data (jika diperlukan)

2. **Eksplorasi Awal**

- Analisis deskriptif statistik
- Analisis korelasi antar variabel
- Visualisasi awal data

3. **Pemodelan Regresi Linear**

- Pembangunan model regresi linear berganda
- Estimasi parameter regresi menggunakan metode kuadrat terkecil (*Ordinary Least Squares*)

4. **Uji Asumsi Model Regresi**

- Uji linearitas (melalui scatterplot residual)
- Uji normalitas residual (plot Q-Q atau uji Shapiro-Wilk)
- Uji homoskedastisitas (plot residual vs fitted)
- Uji multikolinearitas (dengan VIF - Variance Inflation Factor)

5. **Evaluasi Model**

- Menggunakan metrik R-squared dan Adjusted R-squared
- Perhitungan nilai MSE dan RMSE
- Interpretasi signifikansi koefisien (uji t) dan uji model keseluruhan (uji F)

BAB IV

HASIL DAN PEMBAHASAN

4.1 Pra Pemrosesan Data

Sebelum melakukan analisis lebih lanjut dan membangun model regresi linear berganda, dilakukan tahapan pra-pemrosesan data guna memastikan bahwa data yang digunakan bersih, siap analisis, dan memenuhi syarat-syarat statistik dasar.

1. **Pengecekan Data Kosong (Missing Values)**

   Langkah pertama yang dilakukan adalah memeriksa apakah terdapat data yang hilang (missing) pada variabel-variabel yang digunakan dalam pemodelan, yaitu GrLivArea (Luas bangunan yang berada di atas permukaa tanah), GarageArea (Luas garasi rumah) , TotalBsmtSF (Total luas basement), OverallQual (Kualitas keseluruhan rumah), YearBuilt (Tahun dibangun nya rumah), dan SalePrice (Harga jual rumah).

```
> colSums(is.na(modlin_selected))
  GrLivArea  GarageArea TotalBsmtSF OverallQual   YearBuilt   SalePrice
          0           0           0           0           0           0
```

   Berdasarkan hasil pengecekan, tidak ditemukan adanya nilai kosong pada keenam variabel tersebut. Oleh karena itu, tidak diperlukan proses imputasi atau penghapusan data karena semua observasi valid dan lengkap.

2. **Transformasi Data Kategorikal**

```
tibble [1,460 × 6] (S3: tbl_df/tbl/data.frame)
 $ GrLivArea  : num [1:1460] 1710 1262 1786 1717 2198 ...
 $ GarageArea : num [1:1460] 548 460 608 642 836 480 636 484 468 205 ...
 $ TotalBsmtSF: num [1:1460] 856 1262 920 756 1145 ...
 $ OverallQual: num [1:1460] 7 6 7 7 8 5 8 7 7 5 ...
 $ YearBuilt  : num [1:1460] 2003 1976 2001 1915 2000 ...
 $ SalePrice  : num [1:1460] 208500 181500 223500 140000 250000 ...
```

   Semua variabel yang digunakan dalam model bersifat numerik dan berskala rasio atau ordinal, sehingga tidak memerlukan proses transformasi data kategorikal menjadi numerik. Variabel seperti OverallQual memang bersifat ordinal (skala 1 – 10), namun tetap dapat digunakan secara langsung dalam model regresi karena skala nilainya merepresentasikan urutan kualitas dengan makna yang jelas.

3. **Deteksi dan Penanganan Outlier**

Deteksi outlier dilakukan melalui pendekatan visual menggunakan Interquartile Range (IQR).

```
> modlin_selected[modlin_selected$GrLivArea > upper_bound, ]
# A tibble: 31 × 6
   GrLivArea GarageArea TotalBsmtSF OverallQual YearBuilt SalePrice
       <dbl>      <dbl>       <dbl>       <dbl>     <dbl>     <dbl>
 1      2945        641        1410          10      2006    438780
 2      3222        594        1673           7      1990    320000
 3      3608        840        1107          10      1892    475000
 4      3112        795        1360           8      1918    235000
 5      2794        810        1462           8      1995    403000
 6      3493        870        1470           7      1880    295000
 7      2978        564         710           7      1967    242000
 8      3228        546        3200           8      1992    430000
 9      4676        884        3138          10      2007    184750
10      2775        880        1237          10      1893    325000
# i 21 more rows
# i Use `print(n = ...)` to see more rows
```

Hasil eksplorasi awal terhadap variabel-variabel utama dalam model regresi—yakni **GrLivArea**, **GarageArea**, dan **TotalBsmtSF**—menunjukkan adanya sejumlah nilai ekstrem (outlier) yang secara visual tampak berada jauh di atas rentang mayoritas data. Sebagai contoh, pada variabel GrLivArea, terdeteksi lebih dari 30 observasi yang berada jauh di atas batas atas (upper bound) IQR. Namun, setelah dilakukan pemeriksaan lebih lanjut terhadap karakteristik data tersebut, diketahui bahwa nilai-nilai ekstrem tersebut memang mewakili rumah-rumah dengan luas bangunan yang sangat besar, kualitas tinggi, dan harga jual yang sepadan. Hal ini mengindikasikan bahwa meskipun nilai-nilai tersebut tergolong sebagai outlier secara statistik, secara kontekstual dan substansial mereka masih masuk akal dan mencerminkan kondisi nyata di pasar properti.

Oleh karena itu, tidak dilakukan penghapusan atau transformasi data outlier. Keputusan ini diambil agar model regresi yang dibangun tetap mampu mencerminkan keragaman karakteristik rumah yang sebenarnya ada di lapangan, termasuk rumah-rumah dengan spesifikasi premium yang jumlahnya memang lebih sedikit, namun signifikan dalam konteks penetapan harga.

4. **Standardisasi dan Normalisasi Data**

Pada tahap awal analisis, sempat dilakukan eksplorasi terhadap kemungkinan penerapan teknik standardisasi (Z-score) dan normalisasi (Min-Max Scaling) pada variabel numerik. Kedua metode ini umumnya digunakan untuk menyetarakan skala antarvariabel agar tidak ada satu pun variabel yang mendominasi model secara tidak proporsional. Namun, setelah mempertimbangkan bahwa model yang

digunakan adalah regresi linear berganda tanpa regularisasi, dan bahwa seluruh variabel telah berada pada skala yang wajar serta memiliki satuan yang mudah diinterpretasikan secara langsung (seperti meter persegi untuk luas bangunan atau tahun untuk usia bangunan), maka diputuskan untuk tidak menggunakan transformasi skala dalam pemodelan akhir.

Dengan melalui tahapan pra-pemrosesan ini, maka data yang digunakan sudah siap untuk dilakukan eksplorasi awal dan pembangunan model regresi.

4.2 Eksplorasi Awal

1. **Analisis Deskriptif Statistik**

   Analisis deskriptif dilakukan terhadap enam variabel utama. Tujuannya adalah untuk melihat sebaran nilai, kecenderungan pusat (rata-rata, median), serta variasi antar data.

Variabel	Min	Max	Mean	Median
GrLivArea	334	5642	1515	1464
GarageArea	0	1418	473	480
TotalBsmtSF	0	6110	1057.4	991.5
OverallQual	1	10	6.099	6
YearBuilt	1872	2010	1971	1973
SalePrice	34900	755000	180921	163000

   Data deskriptif ini membantu memastikan bahwa data cukup beragam dan tidak terdistribusi secara sempit.

2. **Analisis Korelasi Antar Variabel**

   Korelasi Pearson digunakan untuk mengukur hubungan linier antar variabel. Berikut adalah korelasi antara variabel independen terhadap SalePrice:

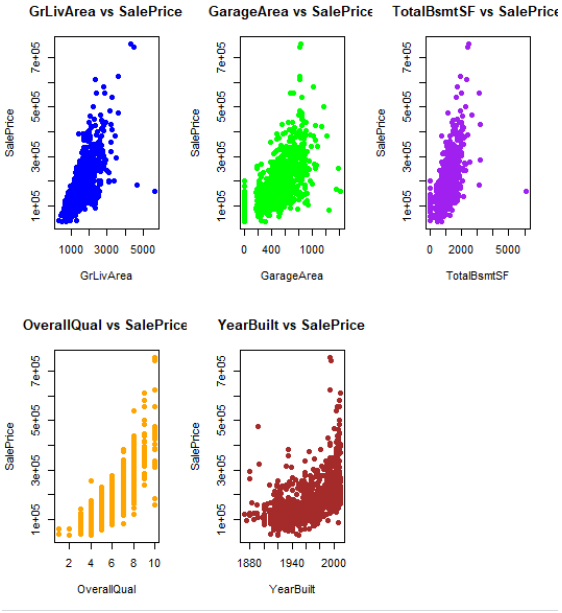
```
> cor(modlin_selected$GrLivArea, modlin_selected$SalePrice)
[1] 0.7086245
> cor(modlin_selected$GarageArea, modlin_selected$SalePrice)
[1] 0.6234314
> cor(modlin_selected$TotalBsmtSF, modlin_selected$SalePrice)
[1] 0.6135806
> cor(modlin_selected$OverallQual, modlin_selected$SalePrice)
[1] 0.7909816
> cor(modlin_selected$YearBuilt, modlin_selected$SalePrice)
[1] 0.5228973
```

   Dari gambar di atas terlihat bahwa OverallQual dan GrLivArea memiliki hubungan paling kuat dengan SalePrice. Hal ini mendukung teori bahwa kualitas dan luas

bangunan sangat memengaruhi nilai jual rumah. Sementara itu, YearBuilt juga memberikan kontribusi positif, meski tidak sekuat variabel lainnya.

3. **Visualisasi Awal Data**

Visualisasi Scatterplot dilakukan untuk melihat pola hubungan antar variabel.



Scatterplot antara GrLivArea, OverallQual, dan TotalBsmtSF terhadap SalePrice menunjukkan pola hubungan linier positif.

Eksplorasi awal menunjukkan bahwa variabel-variabel yang digunakan dalam model memiliki hubungan yang logis dan kuat terhadap harga jual rumah. Pola hubungan linier yang terdeteksi mengindikasikan bahwa regresi linear adalah metode yang sesuai untuk digunakan dalam pemodelan harga rumah ini.

4.3 Pemodelan Regresi Linear

Model regresi linear berganda digunakan untuk memprediksi SalePrice (harga jual rumah) sebagai variabel dependen (Y), berdasarkan lima variabel independen (X).

$$SalePrice = \beta_0 + \beta_1 . GrLivArea + \beta_2 . GarageArea + \beta_3 . TotalBsmtTF + \beta_4 . OverallQual + \beta_5 . YearBuilt$$

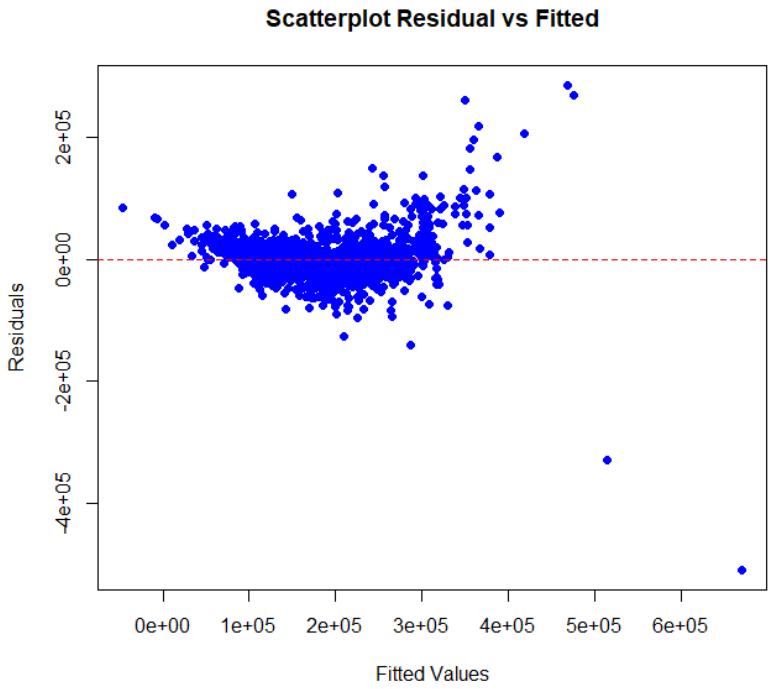
```
Call:
lm(formula = SalePrice ~ GrLivArea + GarageArea + TotalBsmtSF +
    OverallQual + YearBuilt, data = modlin_selected)

Residuals:
    Min      1Q  Median      3Q     Max
-509902  -19599   -2154   15088  286161

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.355e+05  8.310e+04  -8.850  < 2e-16 ***
GrLivArea    5.127e+01  2.567e+00  19.973  < 2e-16 ***
GarageArea   4.598e+01  6.187e+00   7.432 1.82e-13 ***
TotalBsmtSF  2.767e+01  2.874e+00   9.631  < 2e-16 ***
OverallQual  2.099e+04  1.148e+03  18.277  < 2e-16 ***
YearBuilt    3.346e+02  4.369e+01   7.660 3.37e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38440 on 1454 degrees of freedom
Multiple R-squared:  0.7667,    Adjusted R-squared:  0.7659
F-statistic: 955.8 on 5 and 1454 DF,  p-value: < 2.2e-16
```

Berdasarkan hasil estimasi model regresi linear berganda, diperoleh bahwa:

- Koefisien GrLivArea sebesar **51.27** menunjukkan bahwa setiap penambahan 1 m^2 luas bangunan akan meningkatkan harga jual rumah sebesar 51.27 satuan harga, dengan asumsi variabel lain tetap.
- Koefisien GarageArea sebesar **45.98** menunjukkan bahwa setiap penambahan 1 m^2 luas bangunan akan meningkatkan harga jual rumah sebesar 45.98 satuan harga, dengan asumsi variabel lain tetap.
- Koefisien TotalBsmtSF sebesar **27.67** menunjukkan bahwa setiap penambahan 1 m^2 luas bangunan akan meningkatkan harga jual rumah sebesar 27.67 satuan harga, dengan asumsi variabel lain tetap.
- Variabel OverallQual memiliki pengaruh terbesar, yaitu sekitar **20990**, yang berarti rumah dengan kualitas lebih tinggi cenderung memiliki harga jual yang jauh lebih tinggi.
- Koefisien YearBuilt sebesar **334.6** menunjukkan bahwa setiap rumah yang lebih baru 1 tahun dari sebelumnya akan meningkatkan harga jual rumah sebesar 334.6 satuan harga, dengan asumsi variabel lain tetap.
- Semua variabel memiliki hubungan positif dan signifikan terhadap harga jual rumah (p-value < 0.05).

4.4 Uji Asumsi Model Regresi

1. **Uji Linearitas (Scatterplot Residual vs Fitted)**



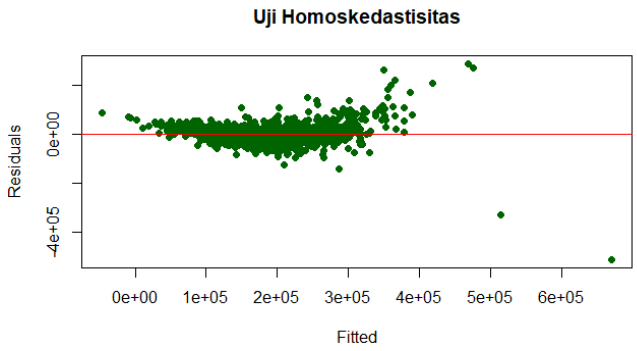
Pada gambar terlihat bahwa pola membentuk melengkung (U). Maka, asumsi linearitas masih belum terpenuhi.

2. Uji Normalitas

```
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.80299, p-value < 2.2e-16
```

Karena p-value lebih kecil daripada 0.05, maka normalitas ditolak. Artinya residual tidak normal.

3. Uji Homoskedastisitas (Residual vs Fitted)



Pada gambar, terlihat bahwa titik residual relative menyebar lebih luas kearah sebelah kanan. Sebaran nya juga tidak sepenuhnya acak dan datar. Maka dari itu asumsi homoskedastisitas nya belum terpenuhi.

4. **Uji Multikolinearitas (VIF)**

```
> vif(model)
  GrLivArea  GarageArea TotalBsmtSF OverallQual   YearBuilt
   1.796600    1.728379    1.569473    2.490327    1.719424
```

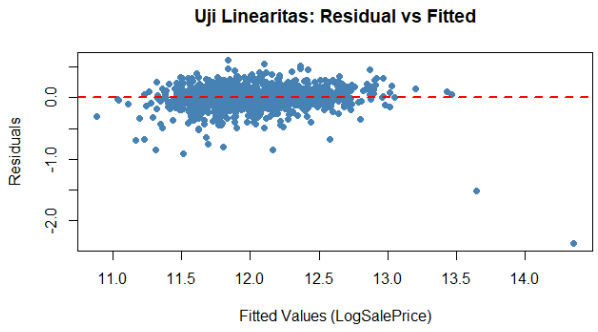
Semua variabel memiliki VIF dibawah 5, maka tidak terjadi tidak terjadi multikolinearitas.

Berdasarkan hasil uji asumsi model regresi awal, ditemukan bahwa residual belum memenuhi asumsi linearitas, normalitas, dan homoskedastisitas. Oleh karena itu, dilakukan transformasi terhadap variabel respon SalePrice menggunakan logaritma natural (log(SalePrice)) untuk memperbaiki sifat distribusinya.

4.5 Uji Asumsi Model Regresi (Setelah Transformasi)

```
> modlin_selected$LogSalePrice <- log(modlin_selected$SalePrice)
>
> model_log <- lm(LogSalePrice ~ GrLivArea + GarageArea + TotalBsmtSF + OverallQu
al + YearBuilt,
+                 data = modlin_selected)
>
> summary(model_log)

Call:
lm(formula = LogSalePrice ~ GrLivArea + GarageArea + TotalBsmtSF +
    OverallQual + YearBuilt, data = modlin_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-2.37003 -0.07735  0.01155  0.09234  0.61374

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.6409021  0.3787851  14.892   <2e-16 ***
GrLivArea   0.0002445  0.0000117  20.897   <2e-16 ***
GarageArea  0.0002581  0.0000282   9.151   <2e-16 ***
TotalBsmtSF 0.0001115  0.0000131   8.512   <2e-16 ***
OverallQual 0.1070406  0.0052334  20.453   <2e-16 ***
YearBuilt   0.0025972  0.0001991  13.043   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1752 on 1454 degrees of freedom
Multiple R-squared:  0.8083,    Adjusted R-squared:  0.8077
F-statistic:  1226 on 5 and 1454 DF,  p-value: < 2.2e-16
```

1. **Uji Linearitas**



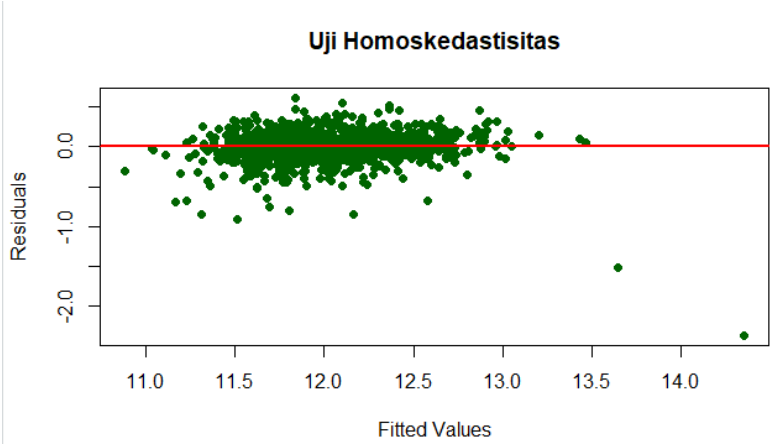
Setelah melakukan transformasi, dapat dilihat dari gambar bahwa tidak ada pola sistemastis melengkung. Maka dari itu asumsi linearitas sudah jauh lebih baik terpenuhi.

2. **Uji Normalitas**

```
> shapiro.test(model_log$residuals)

        Shapiro-Wilk normality test

data:  model_log$residuals
W = 0.85706, p-value < 2.2e-16
```

Setelah dilakukan transformasi, dapat dilihat bahwa p-value lebih kecil daripada 0.05, maka normalitas ditolak. Artinya residual setelah ditransformasi tidak normal.

3. **Uji Homoskedastisitas**



Setelah dilakukan transformasi, terlihat bahwa varians residual tampak konstan di berbagai nilai fitted. Artinya, asumsi homoskedastisitas terpenuhi secara visual.

4. **Uji Multikolinearitas (VIF)**

```
> library(car)
>
> vif(model_log)
  GrLivArea  GarageArea TotalBsmtSF OverallQual   YearBuilt
   1.796600    1.728379    1.569473    2.490327    1.719424
```

Semua variabel memiliki VIF dibawah 5, maka tidak terjadi tidak terjadi multikolinearitas.

Secara keseluruhan, model regresi yang telah dibangun dinilai layak untuk dianalisis dan dievaluasi lebih lanjut, meskipun terdapat pelanggaran ringan pada asumsi normalitas residual.

4.6 Evaluasi Model

Model regresi linear berganda yang telah dibentuk dengan transformasi logaritmik terhadap variabel SalePrice dievaluasi berdasarkan beberapa indikator statistik utama, yaitu koefisien determinasi (R^2 dan Adjusted R^2), nilai MSE dan RMSE, serta hasil uji signifikansi parameter (uji t) dan uji signifikansi model keseluruhan (uji F).

```
> summary(model_log)

Call:
lm(formula = LogSalePrice ~ GrLivArea + GarageArea + TotalBsmtSF +
    OverallQual + YearBuilt, data = modlin_selected)

Residuals:
     Min       1Q   Median       3Q      Max
-2.37003 -0.07735  0.01155  0.09234  0.61374

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.6409021  0.3787851  14.892   <2e-16 ***
GrLivArea   0.0002445  0.0000117  20.897   <2e-16 ***
GarageArea  0.0002581  0.0000282   9.151   <2e-16 ***
TotalBsmtSF 0.0001115  0.0000131   8.512   <2e-16 ***
OverallQual 0.1070406  0.0052334  20.453   <2e-16 ***
YearBuilt   0.0025972  0.0001991  13.043   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1752 on 1454 degrees of freedom
Multiple R-squared:  0.8083,    Adjusted R-squared:  0.8077
F-statistic:  1226 on 5 and 1454 DF,  p-value: < 2.2e-16
```

Nilai R-squared sebesar 80.83% menunjukkan bahwa model dapat menjelaskan sekitar 80.83% variasi harga rumah (dalam skala logaritmik). Nilai Adjusted R-squared yang tinggi menandakan bahwa model tidak overfit dan relevan terhadap data. Pada uji-t terlihat bahwa semua variabel memiliki nilai p-value < 0.05, maka variabel tersebut berpengaruh signifikan secara statistik terhadap harga rumah (dalam skala logaritmik). Pada uji F statistic, terlihat bahwa p-value sangat kecil, maka model secara keseluruhan signifikan secara statistic. Artinya, setidaknya satu variabel prediktor berpengaruh terhadap respon. Selain itu, nilai RMSE sebesar 0.1748 menunjukkan bahwa deviasi rata-rata hasil prediksi terhadap nilai aktual dalam skala log relatif kecil.

Dengan demikian, meskipun asumsi normalitas residual belum sepenuhnya terpenuhi, model ini dapat dikatakan cukup baik dalam menjelaskan variasi dan memprediksi harga jual rumah.

4.7 Strategi Model Building: Uji Interaksi dan Polinomial

Sebagai bagian dari strategi pembangunan model regresi, dilakukan eksplorasi terhadap bentuk model alternatif, yaitu dengan menambahkan unsur interaksi dan polinomial ke dalam model regresi linear berganda.

1. **Model Interaksi**

   Model ini ditulis sebagai:

$$y = \beta_0 + \beta_1(GrLivArea \times OverallQual) + \beta_2.GarageArea + \beta_3.TotalBsmtSF + \beta_4.YearBuilt + \epsilon$$

   Berikut adalah hasil analisis model interaksi:

```
> model_interaksi <- lm(SalePrice ~ GrLivArea * OverallQual + GarageArea + TotalB
smtSF + YearBuilt, data = data)
> summary(model_interaksi)

Call:
lm(formula = SalePrice ~ GrLivArea * OverallQual + GarageArea +
    TotalBsmtSF + YearBuilt, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-617456  -17067   -1463   13426  251654

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -6.732e+05  8.099e+04  -8.311  < 2e-16 ***
GrLivArea           -1.437e+01  7.424e+00  -1.936 0.053087 .
OverallQual          6.704e+03  1.886e+03   3.554 0.000392 ***
GarageArea           4.473e+01  6.012e+00   7.440 1.71e-13 ***
TotalBsmtSF          2.320e+01  2.832e+00   8.193 5.52e-16 ***
YearBuilt            3.524e+02  4.248e+01   8.296 2.42e-16 ***
GrLivArea:OverallQual 9.782e+00  1.042e+00   9.386  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37330 on 1453 degrees of freedom
Multiple R-squared:  0.7801,    Adjusted R-squared:  0.7791
F-statistic: 858.9 on 6 and 1453 DF,  p-value: < 2.2e-16
```

   Hasil pengujian model dengan penambahan interaksi antara GrLivArea dan OverallQual menunjukkan bahwa variabel interaksi tersebut signifikan secara statistik (p-value < 2e-16), dengan nilai koefisien sebesar 9782. Hal ini mengindikasikan bahwa pengaruh luas bangunan terhadap harga rumah dipengaruhi oleh kualitas rumah secara keseluruhan. Model dengan interaksi ini menghasilkan nilai Adjusted R-squared sebesar 0.7791, sedikit lebih tinggi dibandingkan model dasar tanpa interaksi. Oleh karena itu, model dengan interaksi dapat dipertimbangkan sebagai alternatif yang lebih informatif dalam menjelaskan variasi harga rumah.

2. **Model Polinomial**

   Model ini ditulis sebagai:

$$y = \beta_0 + \beta_1.GrLivArea + \beta_2(GrLivArea)^2 + \beta_3.OverallQuil + \beta_4.GarageArea + \beta_5.TotalBsmtSF + \beta_6.YearBuilt + \epsilon$$

Berikut adalah hasil analisis model interaksi:

```
> model_polinomial <- lm(SalePrice ~ GrLivArea + I(GrLivArea^2) + OverallQual + G
arageArea + TotalBsmtSF + YearBuilt, data = data)
> summary(model_polinomial)

Call:
lm(formula = SalePrice ~ GrLivArea + I(GrLivArea^2) + OverallQual +
    GarageArea + TotalBsmtSF + YearBuilt, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-472913  -19838   -2183   15373  302349

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.326e+05  8.306e+04  -8.821  < 2e-16 ***
GrLivArea       6.282e+01  6.951e+00   9.038  < 2e-16 ***
I(GrLivArea^2) -3.029e-03  1.694e-03  -1.788   0.0739 .
OverallQual     2.067e+04  1.161e+03  17.813  < 2e-16 ***
GarageArea      4.556e+01  6.187e+00   7.363 3.00e-13 ***
TotalBsmtSF     2.882e+01  2.942e+00   9.796  < 2e-16 ***
YearBuilt       3.287e+02  4.378e+01   7.509 1.04e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38410 on 1453 degrees of freedom
Multiple R-squared:  0.7672,    Adjusted R-squared:  0.7663
F-statistic: 798.2 on 6 and 1453 DF,  p-value: < 2.2e-16
```

Selain model interaksi, juga diuji model regresi dengan unsur polinomial, yakni penambahan GrLivArea² sebagai variabel kuadrat dari luas bangunan. Hasil analisis menunjukkan bahwa variabel GrLivArea² tidak signifikan secara statistik (p-value = 0.0739), yang berarti tidak terdapat hubungan non-linear yang kuat antara luas bangunan dan harga rumah. Selain itu, nilai Adjusted R-squared dari model polinomial (0.7663) lebih rendah dibandingkan model sebelumnya, serta menghasilkan residual standard error yang lebih tinggi. Berdasarkan hal tersebut, model polinomial dinilai kurang optimal dan tidak digunakan sebagai model akhir.

4.8 Perbandingan Tiga Model

```
> AIC(model_log, model_interaksi, model_polinomial)
                df        AIC
model_log        7  -935.0357
model_interaksi  8 34893.0567
model_polinomial 8 34975.7894
> BIC(model_log, model_interaksi, model_polinomial)
                df        BIC
model_log        7  -898.0324
model_interaksi  8 34935.3463
model_polinomial 8 35018.0789
```

Berdasarkan perbandingan nilai AIC dan BIC, model regresi logaritmik (log(SalePrice)) menunjukkan performa terbaik dengan nilai AIC sebesar -935.04 dan BIC sebesar -898.03. Nilai ini jauh lebih rendah dibandingkan dengan model interaksi (AIC = 34,893.06) dan model polinomial (AIC = 34,975.79), yang menunjukkan bahwa model log

tidak hanya mampu menangkap variasi data secara efisien, tetapi juga memiliki struktur yang lebih sederhana dengan penalti kompleksitas yang rendah. Oleh karena itu, meskipun model interaksi secara statistik signifikan dan memiliki Adjusted R^2 tinggi, model logaritmik tetap menjadi pilihan utama berdasarkan kriteria informasi AIC dan BIC.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis terhadap data harga jual rumah menggunakan model regresi linear berganda, serta melalui serangkaian tahapan pra-pemrosesan, eksplorasi awal, pemodelan, dan evaluasi, diperoleh beberapa kesimpulan sebagai berikut:

1. Variabel-variabel yang berpengaruh signifikan terhadap harga jual rumah (dalam skala logaritmik) adalah:

- GrLivArea (luas bangunan),
- GarageArea (luas garasi),
- TotalBsmtSF (luas basement),
- OverallQual (kualitas keseluruhan rumah), dan
- YearBuilt (tahun dibangun).

2. Kelima variabel ini memiliki nilai signifikansi (p-value) di bawah 0.05 dalam uji-t, yang berarti masing-masing memberikan pengaruh bermakna secara statistik terhadap harga rumah.
3. Model regresi linear berganda yang dibentuk memiliki nilai R-squared dan Adjusted R-squared yang tinggi, menunjukkan bahwa model dapat menjelaskan sebagian besar variasi harga jual rumah. Selain itu, nilai RMSE menunjukkan tingkat kesalahan prediksi yang relatif kecil.
4. Dari hasil uji asumsi, model memenuhi asumsi linearitas, homoskedastisitas, dan tidak adanya multikolinearitas. Namun, asumsi normalitas residual belum sepenuhnya terpenuhi. Meskipun demikian, mengingat ukuran data yang cukup besar, pelanggaran ini tidak terlalu memengaruhi kemampuan model dalam melakukan estimasi dan prediksi.
5. Secara keseluruhan, model regresi yang dibangun layak digunakan untuk memprediksi harga jual rumah, dengan catatan interpretasi dilakukan secara hati-hati.

6. Selain itu, sebagai bagian dari strategi pemodelan lanjutan, telah dilakukan pengujian terhadap dua bentuk alternatif model regresi, yaitu model dengan unsur interaksi dan polinomial. Model interaksi antara variabel GrLivArea dan OverallQual terbukti signifikan secara statistik dan memberikan peningkatan kecil pada Adjusted R-squared, namun tidak dipilih sebagai model akhir karena memiliki nilai AIC dan BIC yang sangat tinggi. Sementara itu, model dengan komponen polinomial $GrLivArea^2$ tidak signifikan dan justru menurunkan performa model secara keseluruhan. Dengan mempertimbangkan seluruh aspek evaluasi model, maka model regresi linear berganda dengan transformasi logaritmik pada variabel SalePrice ditetapkan sebagai model terbaik untuk memprediksi harga rumah dalam studi ini.

5.2 Saran

1. Untuk penelitian selanjutnya, disarankan untuk mempertimbangkan transformasi lanjutan atau penggunaan metode regresi lain seperti regresi robust atau regresi kuantil, terutama jika normalitas residual tetap tidak terpenuhi meski sudah dilakukan transformasi.
2. Perlu dilakukan penambahan variabel seperti kondisi lingkungan, lokasi geografis, serta aksesibilitas fasilitas umum, yang kemungkinan besar juga berpengaruh terhadap harga rumah namun belum dimasukkan dalam model ini.
3. Analisis ini dapat diperluas dengan membandingkan model regresi linear berganda dengan metode prediksi lain, seperti regresi ridge, lasso, atau bahkan model machine learning seperti decision tree dan random forest untuk mendapatkan performa prediksi yang lebih optimal.
4. Selain itu, disarankan untuk melakukan eksplorasi lebih lanjut terhadap model interaksi dan polinomial dengan pendekatan yang lebih kompleks, seperti menggunakan metode non-linear regression, generalized linear models, atau bahkan algoritma machine learning seperti random forest dan gradient boosting. Hal ini dapat membantu menangkap hubungan yang tidak linier dan interaksi yang lebih kompleks antar variabel.

5. Disarankan juga agar pengujian model dilakukan menggunakan teknik validasi silang (cross-validation) untuk memastikan bahwa model tidak hanya cocok terhadap data pelatihan, tetapi juga mampu melakukan prediksi yang baik terhadap data baru. Penggunaan metrik tambahan seperti MAE, MAPE, dan RSME pada data uji dapat memberikan gambaran performa model secara lebih menyeluruh.

DAFTAR PUSTAKA

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

2. Mendenhall, W., & Sincich, T. (2012). *A second course in statistics: Regression Analysis* (7th ed.). Pearson Education.