UIN Jakarta

Tugas Akhir Mata Kuliah Model Linier

PREDIKSI HARGA JUAL RUMAH

MENGGUNAKAN MODEL REGRESI

Kelompok B4

UIN Jakarta

ANGGOTA KELOMPOK:

ALIFIA INTAN	11230940000016
FAIRUUZZARI RAMADHAN	11230940000040
APRILIA NUR AFIFAH	11230940000060
SEKAR AFIFA CETTASTAMI	11230940000062

DAFTAR ISI

LATAR BELAKANG MASALAH

Mengapa prediksi harga rumah penting?

- Perumahan = kebutuhan dasar → permintaan tinggi
- Harga rumah sangat dipengaruhi banyak faktor kompleks
- Dibutuhkan model statistik untuk analisis objektif

Solusi:

Regresi linear → sederhana, mudah diinterpretasi, cocok untuk prediksi

RUMUSAN MASALAH

Rumusan Masalah

- Faktor apa saja yang memengaruhi harga jual rumah?
- Seberapa baik model regresi linear memprediksi harga?

Tujuan

- Mengidentifikasi variabel-variabel signifikan
- Membangun model regresi linear yang akurat dan interpretable

DATASET & VARIABEL PENELITIAN

Penelitian ini menggunakan data sekunder dari dataset harga rumah dalam format spreadsheet. Data mencakup karakteristik properti seperti luas tanah, tipe bangunan, zona pemukiman, jumlah lantai, dan lainnya yang memengaruhi harga jual rumah. Semua variabel bersifat kuantitatif dan siap untuk dianalisis.

Variabel Respon (Y):

- SalePrice – Harga jual rumah

Variabel Prediktor (X):

- GrLivArea – Luas bangunan atas tanah
- GarageArea – Luas garasi
- TotalBsmtSF – Luas basement
- OverallQual – Kualitas keseluruhan rumah (skala ordinal 1–10)
- YearBuilt – Tahun dibangunnya rumah

METODE ANALISIS: REGRESI LINEAR BERGANDA

Digunakan untuk menganalisis pengaruh beberapa variabel independen terhadap harga jual rumah (variabel dependen).

Pra-pemrosesan Data:

- Cek dan tangani missing values
- Transformasi data kategorikal (jika perlu)
- Deteksi & evaluasi outlier
- Normalisasi atau standardisasi (jika perlu)

Eksplorasi Awal:

- Statistik deskriptif
- Korelasi antar variabel
- Visualisasi awal (scatterplot, dll.)

Pemodelan Regresi:

- Model regresi linear berganda
- Estimasi parameter (Ordinary Least Squares)

Uji Asumsi Model:

- Linearitas (residual vs fitted)
- Normalitas (Q-Q plot, Shapiro-Wilk)
- Homoskedastisitas
- Multikolinearitas (VIF)

Evaluasi Model:

- R-squared & Adjusted R-squared
- MSE & RMSE
- Uji t & Uji F

HASIL DAN PEMBAHASAN

PRA PEMROSESAN DATA

➔ Pengecekan Missing Value

Semua variabel utama (GrLivArea, GarageArea, TotalBsmtSF, OverallQual, YearBuilt, SalePrice) lengkap, tidak ada nilai kosong.

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

Transformasi Data Kategorikal

Semua variabel yang digunakan dalam model bersifat numerik dan berskala rasio atau ordinal, sehingga tidak memerlukan proses transformasi data kategorikal menjadi numerik

```
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 ...
```

➔ Outllier

Deteksi outlier dilakukan melalui pendekatan visual menggunakan Interquartile Range (IQR).
Ada outlier (nilai ekstrem) — yakni **GrLivArea**, **GarageArea**, dan **TotalBsmtSF**, tapi tidak dihapus karena mencerminkan kondisi rumah premium yang valid.

```
> 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
```

Standardisasi dan Normalisasi Data

Tidak dilakukan karena semua variabel berada dalam skala interpretatif yang jelas.

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

EKSPLORASI AWAL

➔ Analisis Deskriptif Statistik

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.

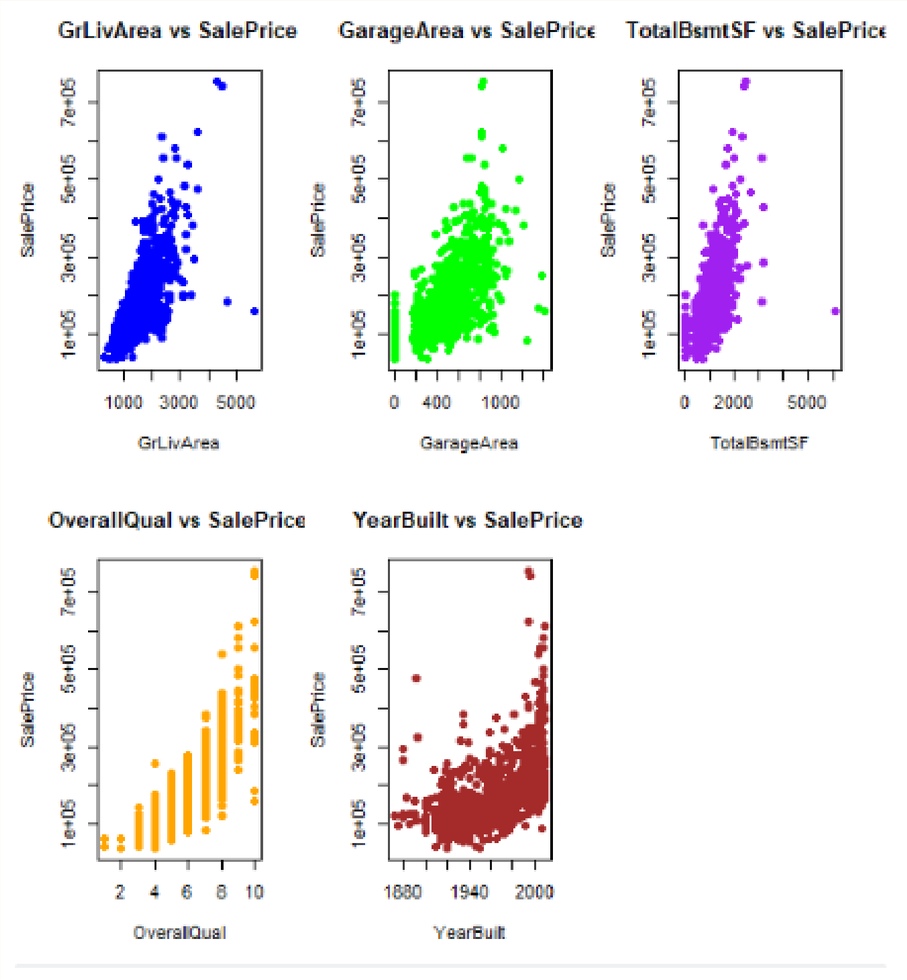
➔ Analisis Korelasi Antar Variabel

```
> 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
```

OverallQual & GrLivArea paling kuat korelasinya dengan SalePrice.

Hal ini mendukung teori bahwa kualitas dan luas bangunan sangat memengaruhi nilai jual rumah.

➔ Visualisasi Awal Data



Scatterplot menunjukkan hubungan positif linier antara variabel-variabel tersebut dengan SalePrice.

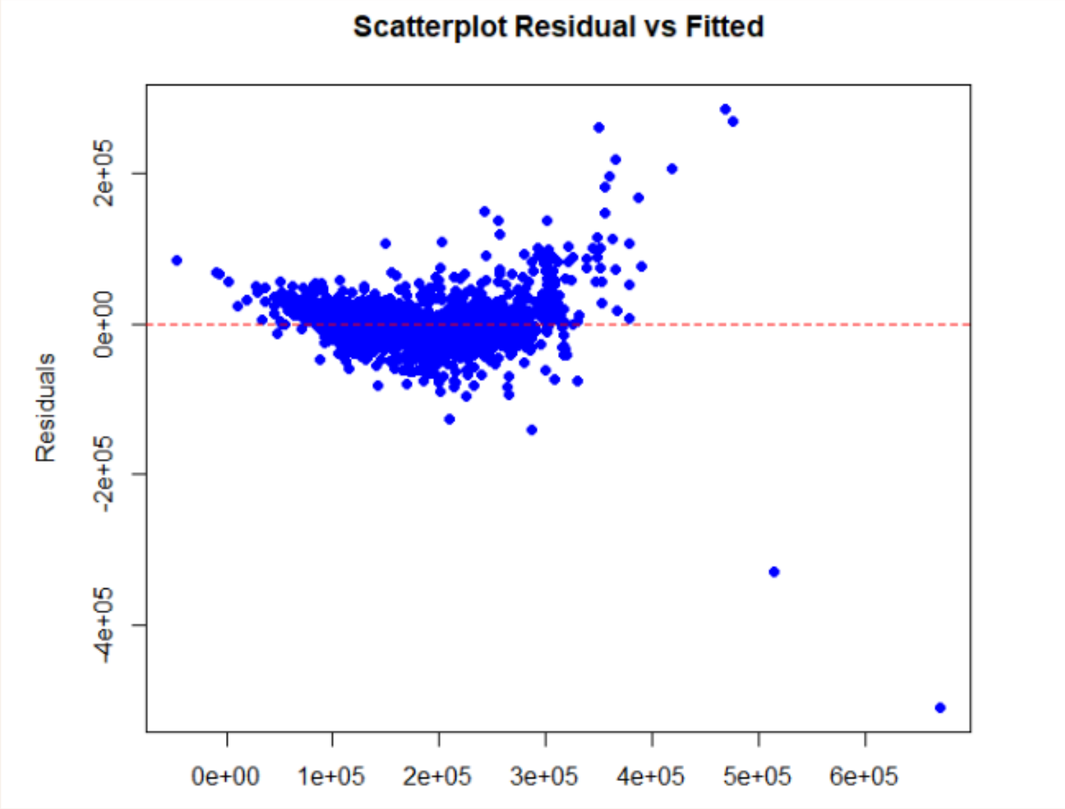
Pola hubungan linier yang terdeteksi mengindikasikan bahwa regresi linear adalah metode yang sesuai untuk digunakan dalam pemodelan harga rumah ini.

PEMODELAN REGRESI LINEAR

```
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
```

- Variabel signifikan: GrLivArea, GarageArea, TotalBsmtSF, OverallQual, YearBuilt.
- Semua koefisien positif.
  - Koefisien GrLivArea sebesar **51.27**
  - Koefisien GarageArea sebesar **45.98**
  - Koefisien TotalBsmtSF sebesar **27.67**
  - OverallQual memiliki pengaruh terbesar, yaitu sekitar **20990**
  - Koefisien YearBuilt sebesar **334.6**
- Semua variabel memiliki hubungan positif dan signifikan terhadap harga jual rumah (p-value < 0.05).

UJI ASUMSI MODEL REGRESI (SEBELUM TRANSFORMASI)

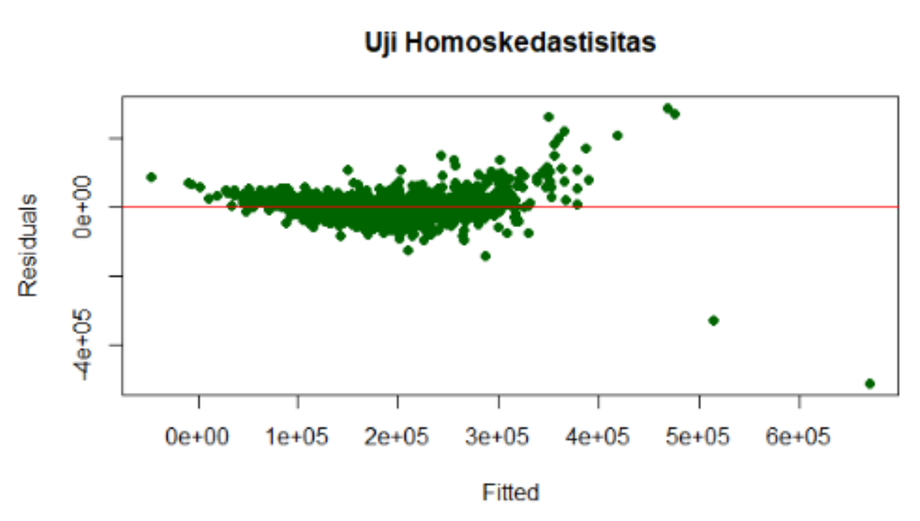


Linearitas ❌: Terlihat pola melengkung (U)

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

        Shapiro-Wilk normality test

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

Normalitas ❌: p-value < 0.05. Artinya residual tidak normal

UJI ASUMSI MODEL REGRESI (SEBELUM TRANSFORMASI)



Homoskedastisitas ❌: Varians residual tidak konstan

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

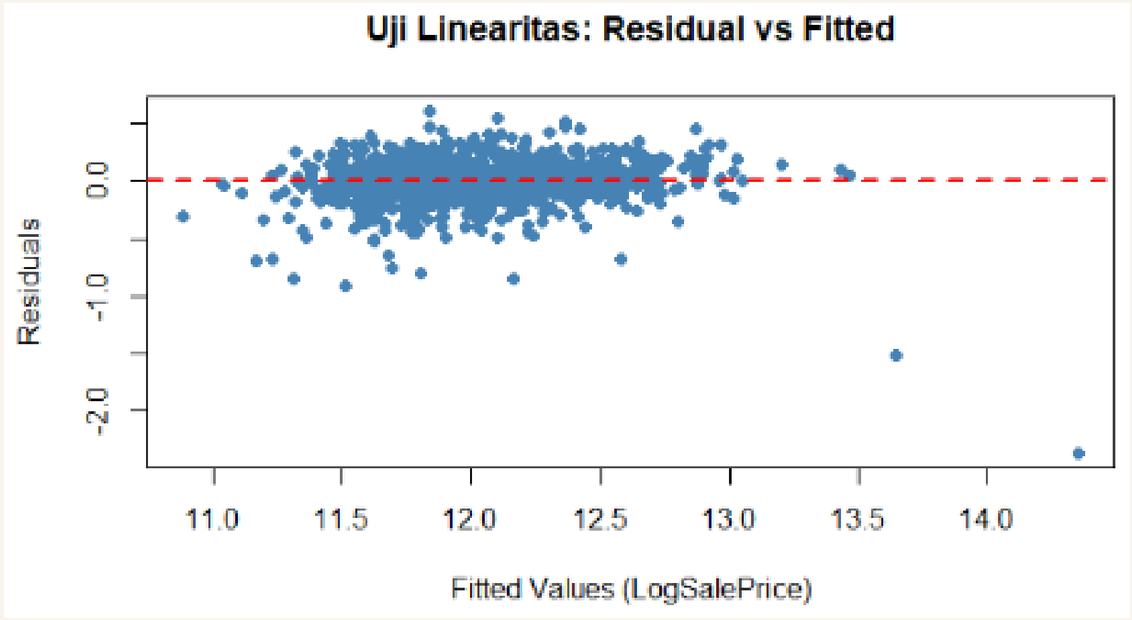
Multikolinearitas ✅: VIF < 5.

Dilakukan transformasi log(SalePrice)

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
```

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.

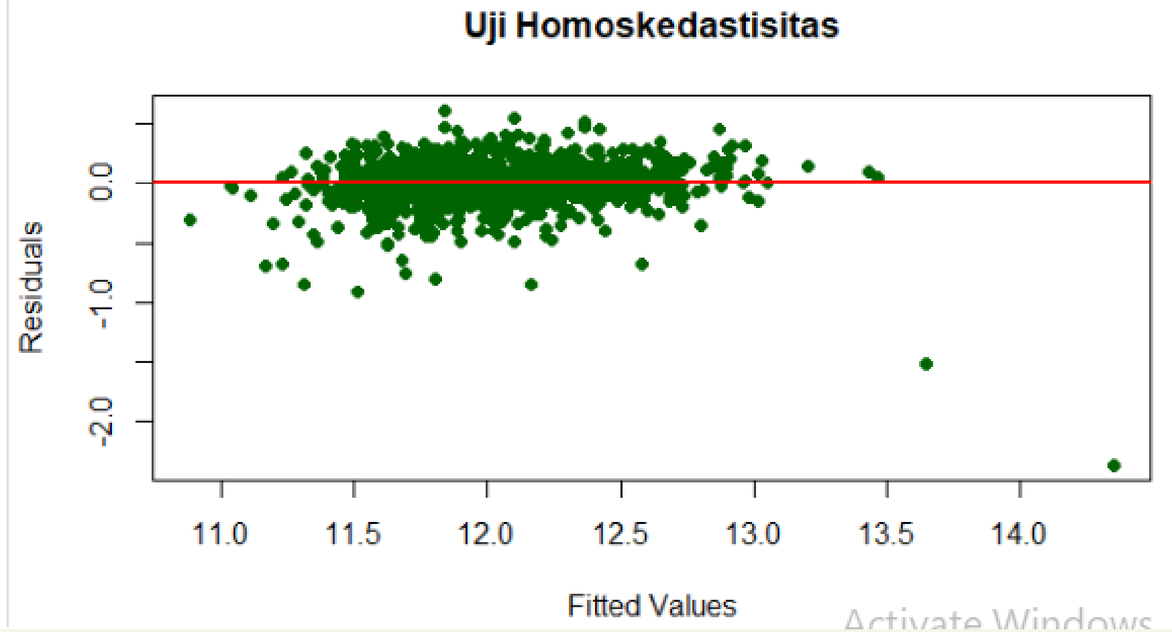
UJI ASUMSI MODEL REGRESI (SETELAH TRANSFORMASI)

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.

Uji Homoskedastisitas



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

UJI ASUMSI MODEL REGRESI (SETELAH TRANSFORMASI)

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.

EVALUASI MODEL

```
> 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
```

- **R^2 = 80.83%** → model menjelaskan variasi dengan baik.
- **RMSE = 0.1748** → kesalahan prediksi relatif kecil.
- **Uji t & Uji F** → semua signifikan ($p < 0.05$)
- Model dianggap **baik & layak** digunakan untuk prediksi.

STRATEGI MODEL BUILDING: Uji Interaksi dan Polinomial

➔ Model Interaksi (GrLivArea * OverallQual)

```
> 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
```

Signifikan, Adjusted $R^2 = 0.7791$

➔ Model Polinomial ($GrLivArea^2$)

```
> 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
```

Tidak signifikan, Adjusted R^2 lebih rendah.

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
```

Model Log(SalePrice)

- Nilai AIC = -935.04 dan BIC = -898.03 menunjukkan bahwa model ini paling efisien dalam menjelaskan data dengan kompleksitas yang minimal.
- Sederhana tapi akurat
- Asumsi regresi paling banyak terpenuhi

Model Interaksi (GrLivArea * OverallQual)

- Adjusted R^2 sedikit naik, tapi tidak terlalu besar dibandingkan model log (hanya sekitar 77.91%).
- Nilai AIC sangat tinggi (34,893.06)
- Bisa dipakai jika ingin interpretasi lebih mendalam, tapi kurang efisien sebagai model prediksi.

Model Polinomial (GrLivArea^2)

- Variabel polinomial tidak signifikan secara statistik (p-value > 0.05).
- Adjusted R^2 menurun dan error meningkat
- AIC = 34,975.79 (lebih buruk dari model interaksi dan log)

KESIMPULAN

Faktor apa saja yang memengaruhi harga jual rumah?

Faktor yang signifikan mempengaruhi harga jual rumah antara lain luas bangunan , garasi , basement , kualitas rumah (OverallQual) , dan tahun pembangunan . Semua variabel ini terbukti berpengaruh secara statistik, dengan OverallQual sebagai yang paling dominan. Temuan ini pentingnya aspek fisik rumah dalam menentukan harga properti dan dapat menjadi acuan bagi pihak terkait.

Bagaimana cara mengidentifikasi variabel yang signifikan dan membangun model regresi linier yang akurat serta dapat diinterpretasikan dengan baik?

Penelitian ini mengidentifikasi variabel yang signifikan berdasarkan nilai p-value (< 0,05), yang menunjukkan pengaruh nyata terhadap harga jual rumah. Seleksi variabel dilakukan untuk mencegah multikolinearitas dan overfitting. Model evaluasi menggunakan nilai R^2, di mana nilai yang tinggi menandakan kemampuan prediktif yang baik serta hubungan antarvariabel yang jelas.

Seberapa efektif model regresi linier dalam menjelaskan dan memprediksi variasi harga berdasarkan variabel-variabel yang tersedia?

Model regresi linier dalam penelitian ini cukup efektif untuk menjelaskan dan memprediksi harga jual rumah. Dengan R-squared sebesar 80,83% dan RMSE yang rendah , model mampu menggambarkan sebagian besar variasi harga dengan tingkat kesalahan prediksi yang kecil. Meski asumsi normalitas sisa belum sepenuhnya terpenuhi, hal ini tidak berdampak besar karena asumsi lainnya telah terpenuhi dan data yang digunakan cukup besar. Secara keseluruhan, model ini layak digunakan untuk analisis dan prediksi harga properti.

SEKIAN & TERIMA KASIH

SEMOGA HAL YANG DIPAPARKAN DAPAT BERMANFAAT BAGI PENULIS DAN PEMBACA!