PROJECT

Memodelkan Hubungan Antara Luas Rumah, Biaya Registrasi, dan Lokasi Rumah dengan Harga Rumah di Kota Chennai, India

Disusun Oleh :

Adawia Ananda	2106724883
Divaya Syifa Susilobudi	2106650790
Nadia Sukesi Sianipar	2106700776
Rafly Witjaksana Hartantyo	2106651572
Rima Fitrianti Azahra	2106701974

Mata Kuliah

Model Linear

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

DEPOK

2023

Tabel Kontribusi Anggota Kelompok:

No	Nama	NPM	Kontribusi	Tingkat Kontribusi
1	Adawia Ananda	2106724883	Terlibat aktif diskusi, mencari dataset, membuat visualisasi.	100%
2	Divaya Syifa Susilobudi	2106650790	Terlibat aktif diskusi, membuat pendahuluan dan penutup, menyiapkan desain ppt	100%
3	Nadia Sukesi Sianipar	2106700776	Aktif dalam diskusi, melakukan pengolahan data dan analisis hasil	100%
4	Rafly Witjaksana Hartantyo	2106651572	Aktif dalam diskusi dan melakukan proses pre-processing	100%
5	Rima Fitrianti Azahra	2106701974	Aktif dalam diskusi, membuat pemodelan	100%

Daftar Isi

Bab 1
Pendahuluan

1.1. Masalah

India sebagai negara yang menempati urutan kedua dengan penduduk terbanyak di dunia harus menghadapi berbagai permasalahan kependudukan di negaranya. Salah satu dari permasalahan tersebut adalah tingkat pertumbuhan penduduk yang tinggi. Kepadatan penduduk ini juga berdampak pada permasalahan perumahan di India. Dengan lahan yang semakin langka, harga perumahan di India pun mengalami peningkatan. Oleh karena itu, para pembeli rumah harus bijak dalam membeli rumah dan memastikan bahwa apa yang ia dapatkan sesuai dengan biaya yang dikeluarkan.

Pada *project* ini, kami akan membuat sebuah model untuk memprediksi harga jual sebuah rumah berdasarkan variabel luas bangunan, biaya registrasi, serta lokasi rumah. Setelah dibuat model yang sesuai, diharapkan calon pembeli dapat membandingkan apakah harga asli rumah merupakan penawaran yang terbaik atau bukan berdasarkan karakteristik yang dimiliki rumah tersebut.

1.2. Deskripsi Dataset

Kami menggunakan dataset *Chennai Housing Sales Price* yang didapatkan dari situs Kaggle. Dataset ini memiliki 7109 baris dan 22 kolom. Untuk kepentingan *project* ini, kami hanya memilih 4 variabel, yaitu sebagai berikut:

Nama	**Tipe Data**	**Unit Pengukuran**	**Deskripsi**
SALES_PRICE	Kuantitatif (Respon)	US Dollar	Harga jual rumah
AREA	Kategorik (Prediktor)	-	Di area mana rumah tersebut berlokasi
INT_SQFT	Kuantitatif (Prediktor)	Kaki persegi	Luas bangunan
REG_FEE	Kuantitatif (Prediktor)	US Dollar	Biaya registrasi setelah penjualan

Bab 2

Pre-processing dan Analisis Deskriptif

2.1 Pre-Processing

Pre-processing data dilakukan untuk mengatasi outlier pada data, menyeleksi variabel-variabel yang diinginkan, serta melihat keterkaitan antar variabel. Pada proses ini, langkah-langkah yang kami lakukan adalah sebagai berikut:

1. Pertama, kami mengecek data type masing masing variable.
2. Kami mencari missing values dari dataset yang kami gunakan.
3. Kami menghapus missing values dari dataset yang kami gunakan.
4. Kami melakukan pengecekan unique values untuk masing masing variable categorik
5. Kami merename value yang *typo* dengan value yang sesuai
6. Kami mengambil 300 sampel data random dari seluruh dataset
7. Kami menghapus variable selain AREA, INT_SQFT, REG_FEE, dan SALES_PRICE
8. Kami mencari outlier menggunakan boxplot dan menghapus outlier tersebut
9. Terakhir, kami menyimpan dataset yang telah dilakukan pre-processing untuk dilakukan pemodelan.

2.2 Visualisasi

2.2.1. Statistik dari Dataset

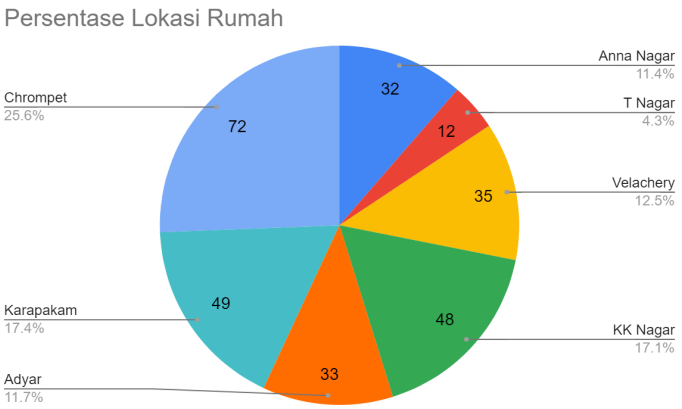
	AREA	INT_SQFT	REG_FEE	SALES_PRICE
count	281	281.000000	281.000000	2.810000e+02
unique	7	NaN	NaN	NaN
top	Chrompet	NaN	NaN	NaN
freq	72	NaN	NaN	NaN
mean	NaN	1359.693950	367992.149466	1.066606e+07
std	NaN	464.843421	123569.183482	3.241106e+06
min	NaN	517.000000	134508.000000	3.785500e+06
25%	NaN	1008.000000	271920.000000	8.371810e+06
50%	NaN	1287.000000	344772.000000	1.026245e+07
75%	NaN	1707.000000	444970.000000	1.274669e+07
max	NaN	2483.000000	695964.000000	1.897681e+07

Interpretasi:

Berdasarkan ringkasan dari statistik yang tertera di atas, dapat diambil beberapa kesimpulan sebagai berikut:

1. Dataset tersebut terdiri dari 7 area berbeda untuk lokasi rumah, dengan Chrompet menjadi area dengan jumlah sampel rumah terbanyak sejumlah 72 rumah.
2. Dataset tersebut memiliki rata-rata luas rumah sebesar 1359 *square feet*, dengan nilai minimal 517 *square feet* dan nilai maximum 2483 *square feet*.
3. Dataset tersebut memiliki rata-rata biaya registrasi sebesar 367992 USD, dengan nilai minimal 134508 USD dan nilai maximum 695964 USD.
4. Dataset tersebut memiliki rata-rata harga penjualan 10666060 USD, dengan nilai minimal 3785500 USD dan nilai maximum 18976810 USD.

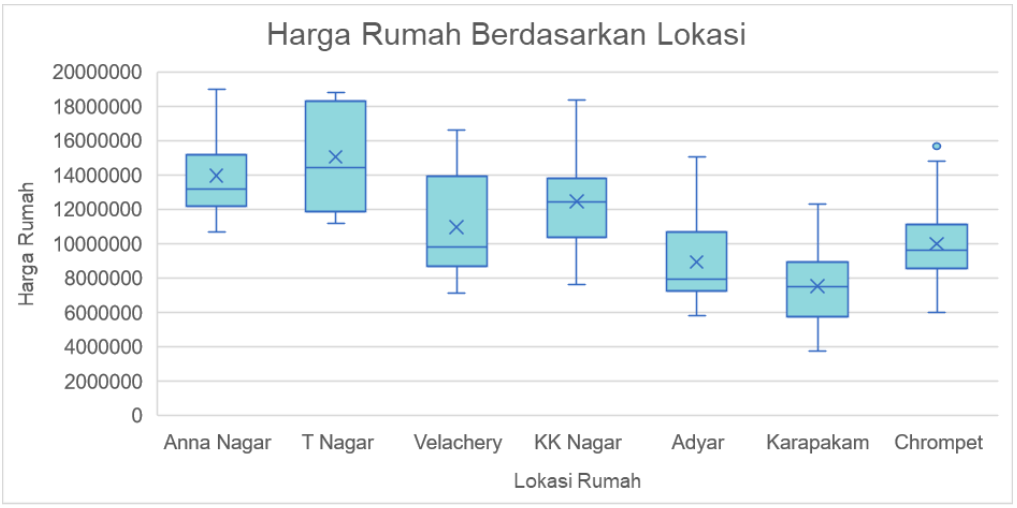
2.2.2. *Pie Chart* Variabel Lokasi



Interpretasi:

Berdasarkan *Pie Chart*, dapat dilihat bahwa Chrompet menjadi area dengan jumlah sampel rumah terbanyak sejumlah 72 rumah. Lalu, diikuti dengan Karapakam dengan 49 rumah dan KK Nagar dengan 48 rumah. Area dengan jumlah sampel rumah paling sedikit adalah T Nagar, dengan 12 rumah.

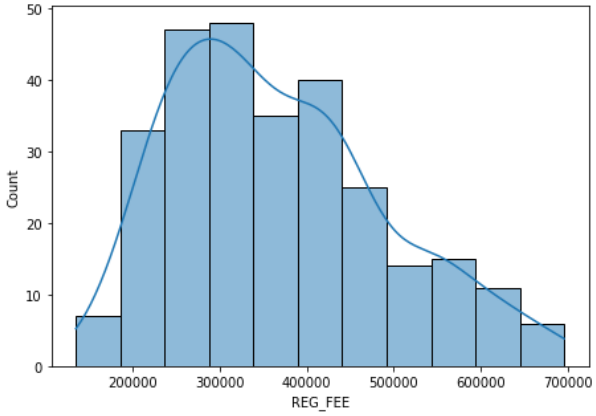
2.2.3. *Box Plot* Harga Rumah Berdasarkan Lokasi



Interpretasi:

Dari ketujuh lokasi, didapati bahwa Karapakam memiliki median harga rumah terendah. Di sisi lain, didapati bahwa T Nagar memiliki median harga rumah tertinggi. Selain itu, terlihat bahwa terdapat outlier pada lokasi Chrompet.

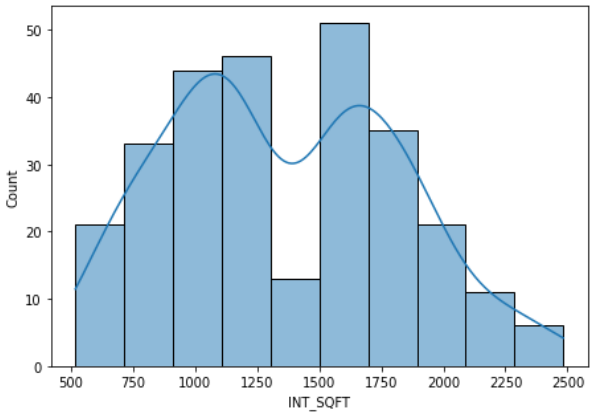
2.2.4. Histogram Variabel Biaya Registrasi



Interpretasi:

Berdasarkan histogram di atas, didapatkan bahwa persebaran frekuensi dari biaya registrasi mendekati distribusi normal dengan frekuensi tertinggi berada di rentang 20000-300000 USD.

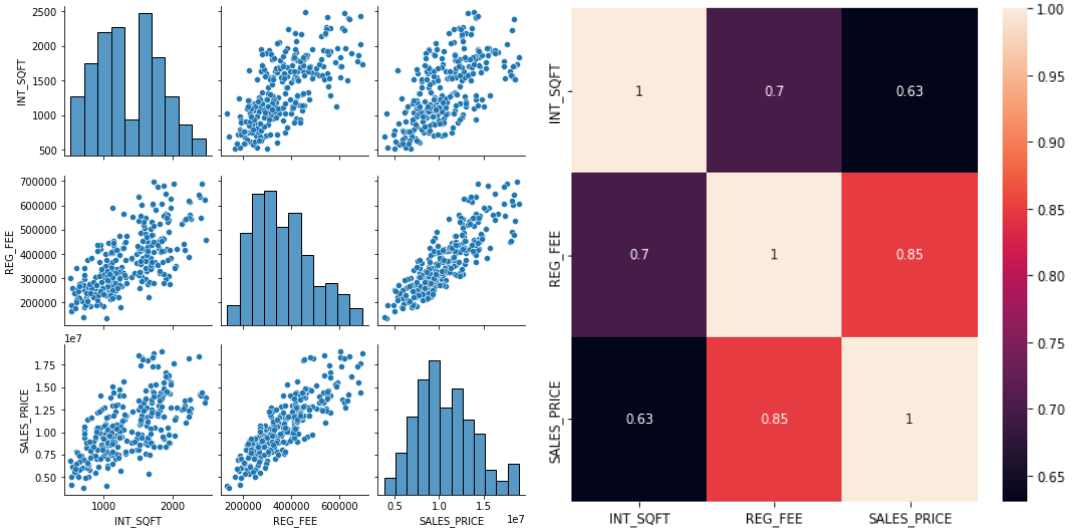
2.2.5. Histogram Variabel Luas Rumah



Interpretasi:

Berdasarkan histogram di atas, didapatkan bahwa persebaran frekuensi dari luas rumah cukup bervariasi. Frekuensi luas rumah paling tinggi terletak pada rentang 1500-1750 *square feet*, dan frekuensi luas rumah yang cenderung rendah terletak pada rentang lebih dari 2250 *square feet*.

2.2.6. Pairplot dan Heatmap antar Variabel



Interpretasi:

Berdasarkan pairplot dan heatmap di atas, didapatkan beberapa kesimpulan sebagai berikut:

1. Terdapat korelasi positif (0.63) pada hubungan variabel dependen SALES_PRICE dengan variabel prediktor INT_SQFT. Didapati pula korelasi positif yang cukup tinggi (0.85) pada hubungan variabel dependen SALES_PRICE dengan variabel prediktof REG_FEE. Oleh karena itu, kedua prediktor tersebut dapat dipertimbangkan sebagai variabel prediktor yang dapat digunakan untuk membuat model regresi linier.
2. Terdapat korelasi positif yang cukup besar (0.7) pada hubungan variabel prediktor REG_FEE dan variabel prediktor (INT_SQFT). Korelasi yang cukup besar ini harus diperhatikan agar tidak terjadi multikolinearitas pada model.

Bab 3
Pemodelan

3.1. Kriteria Model Terbaik

Dalam menentukan model terbaik, kriteria ideal yang sebaiknya dipenuhi adalah sebagai berikut:

1. Memenuhi prinsip *parsimony*, yaitu model berbentuk sederhana dengan parameter yang lebih sedikit lebih disukai daripada model yang kompleks dengan parameter yang lebih banyak, apabila kedua model memiliki *fit* yang cukup sama.
2. Model memenuhi asumsi normalitas, homoskedasitas, linearitas, dan nonmultikolinearitas.
3. Nilai R^2 minimal 85%
4. Nilai VIF maksimal 4
5. Prediktor signifikan untuk alpha = 0.05 dan uji F memberikan hasil penolakan hipotesis null.

3.2. Model Pertama

Model yang ingin kami buat adalah model untuk memprediksi harga rumah (*Sales Price)* berdasarkan luas rumah (INT SQFT), biaya registrasi rumah (REG FEE), dan lokasi rumah (Area) yang disertai dengan interaksi dari tiap-tiap variabel.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{3A} + \ldots + \beta_8 X_{3F} + \beta_9 X_1 X_2 + \ldots + \beta_{26} X_1 X_2 X_{3F} + \epsilon$$

Dengan

Y : Harga Rumah (Sales Price)
X_1 : Luas Rumah (INT SQFT)
X_2 : Biaya Registrasi Rumah (REG FEE)
X_{3A} : Rumah yang berlokasi di Anna Nagar
X_{3B} : Rumah yang berlokasi di Chrompet
X_{3C} : Rumah yang berlokasi di Karakapam
X_{3D} : Rumah yang berlokasi di KK Nagar
X_{3E} : Rumah yang berlokasi di T Nagar
X_{3F} : Rumah yang berlokasi di Velachery
ϵ : error yang bersifat $NIID(0, \sigma^2)$

Output Model:

```
Call:
lm(formula = SALES_PRICE ~ factor(AREA) * INT_SQFT * REG_FEE,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2924492  -756036  -154844   608144  4352556

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              5.028e+05  3.751e+06   0.134   0.8935
factor(AREA)Anna Nagar                  -2.815e+07  1.740e+07  -1.618   0.1069
factor(AREA)Chrompet                     1.772e+06  5.338e+06   0.332   0.7402
factor(AREA)Karapakam                    7.793e+05  4.245e+06   0.184   0.8545
factor(AREA)KK Nagar                    -6.655e+06  5.738e+06  -1.160   0.2472
factor(AREA)T Nagar                     -3.908e+07  2.492e+07  -1.568   0.1181
factor(AREA)Velachery                    1.822e+06  1.021e+07   0.178   0.8586
INT_SQFT                                 1.753e+03  3.532e+03   0.496   0.6200
REG_FEE                                  2.364e+01  1.337e+01   1.769   0.0781 .
factor(AREA)Anna Nagar:INT_SQFT          1.749e+04  1.020e+04   1.715   0.0876 .
factor(AREA)Chrompet:INT_SQFT           -1.743e+03  5.158e+03  -0.338   0.7358
factor(AREA)Karapakam:INT_SQFT          -1.356e+03  3.991e+03  -0.340   0.7343
factor(AREA)KK Nagar:INT_SQFT            4.039e+03  4.223e+03   0.956   0.3398
factor(AREA)T Nagar:INT_SQFT             2.351e+04  1.530e+04   1.537   0.1255
factor(AREA)Velachery:INT_SQFT          -2.471e+03  6.458e+03  -0.383   0.7023
factor(AREA)Anna Nagar:REG_FEE           6.085e+01  4.128e+01   1.474   0.1417
factor(AREA)Chrompet:REG_FEE            -1.443e+00  1.869e+01  -0.077   0.9385
factor(AREA)Karapakam:REG_FEE           -5.459e+00  1.524e+01  -0.358   0.7204
factor(AREA)KK Nagar:REG_FEE             1.065e+01  1.616e+01   0.659   0.5104
factor(AREA)T Nagar:REG_FEE              1.037e+02  5.154e+01   2.011   0.0453 *
factor(AREA)Velachery:REG_FEE           -7.471e+00  2.797e+01  -0.267   0.7896
INT_SQFT:REG_FEE                        -2.134e-03  1.198e-02  -0.178   0.8588
factor(AREA)Anna Nagar:INT_SQFT:REG_FEE -3.592e-02  2.500e-02  -1.437   0.1520
factor(AREA)Chrompet:INT_SQFT:REG_FEE    6.535e-03  1.733e-02   0.377   0.7065
factor(AREA)Karapakam:INT_SQFT:REG_FEE   3.104e-03  1.345e-02   0.231   0.8177
factor(AREA)KK Nagar:INT_SQFT:REG_FEE   -7.525e-03  1.286e-02  -0.585   0.5589
factor(AREA)T Nagar:INT_SQFT:REG_FEE    -6.004e-02  3.235e-02  -1.856   0.0646 .
factor(AREA)Velachery:INT_SQFT:REG_FEE   7.514e-03  1.821e-02   0.413   0.6803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1192000 on 253 degrees of freedom
Multiple R-squared:  0.8778,    Adjusted R-squared:  0.8648
F-statistic: 67.31 on 27 and 253 DF,  p-value: < 2.2e-16
```

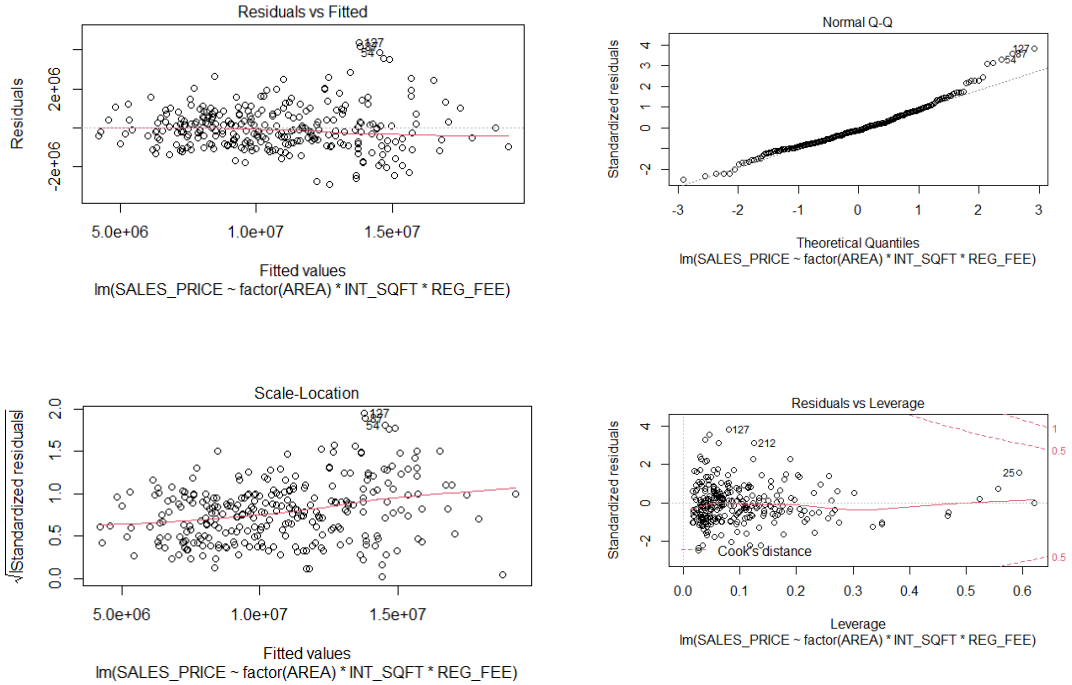
Terlihat dari model tersebut didapat nilai R squared yang cukup besar, yaitu 0,8778. Hal ini menunjukkan bahwa 87% harga rumah dapat diprediksi dengan luas rumah, biaya registrasi rumah, lokasi rumah, beserta interaksinya. Namun, nilai R squared yang besar ini juga bisa disebabkan karena banyaknya variabel prediktor yang digunakan. Jika kita melihat variabel secara univariat, tidak ada yang berpengaruh signifikan terhadap prediksi harga rumah sehingga model ini kurang baik untuk digunakan.

VIF Model:

```
                                    GVIF Df GVIF^(1/(2*Df))
factor(AREA)                3.592419e+18  6        35.17890
INT_SQFT                    5.313757e+02  1        23.05159
REG_FEE                     5.376517e+02  1        23.18732
factor(AREA):INT_SQFT       6.349375e+18  6        36.88878
factor(AREA):REG_FEE        1.207898e+19  6        38.91967
INT_SQFT:REG_FEE            3.119286e+03  1        55.85057
factor(AREA):INT_SQFT:REG_FEE 2.765262e+19  6        41.70083
```

Terlihat pada bagian $GVIF^{(1/(2\times df))}$ nilai yang didapat untuk semua variabel prediktor beserta interaksinya bernilai > 4, yang menandakan model ini belum masuk ke kriteria model terbaik.

Plot Model:



Terlihat bahwa model memenuhi asumsi normalitas, linearitas, dan homoskedasitas.

3.3. Model Kedua (*Final Model*)

Model yang ingin kami buat adalah model untuk memprediksi harga rumah (*Sales Price)* berdasarkan luas rumah (INT SQFT), biaya registrasi rumah (REG FEE), dan lokasi rumah (Area) tanpa disertai dengan interaksi.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{3A} + \ldots + \beta_8 X_{3F} + \epsilon$$

Dengan

Y : Harga Rumah (Sales Price)
X_1 : Luas Rumah (INT SQFT)
X_2 : Biaya Registrasi Rumah (REG FEE)
X_{3A} : Rumah yang berlokasi di Anna Nagar
X_{3B} : Rumah yang berlokasi di Chrompet
X_{3C} : Rumah yang berlokasi di Karakapam
X_{3D} : Rumah yang berlokasi di KK Nagar
X_{3E} : Rumah yang berlokasi di T Nagar
X_{3F} : Rumah yang berlokasi di Velachery
ϵ : error yang bersifat $NIID(0, \sigma^2)$

Output Model:

```
Call:
lm(formula = SALES_PRICE ~ factor(AREA) + INT_SQFT + REG_FEE,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2750696  -823276   -67225   705203  4232509

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.627e+06  3.938e+05   4.130 4.83e-05 ***
factor(AREA)Anna Nagar   1.565e+06  3.910e+05   4.002 8.11e-05 ***
factor(AREA)Chrompet     1.549e+06  2.669e+05   5.802 1.82e-08 ***
factor(AREA)Karapakam   -1.250e+06  2.850e+05  -4.385 1.66e-05 ***
factor(AREA)KK Nagar    -8.707e+05  4.085e+05  -2.131   0.0340 *
factor(AREA)T Nagar      1.906e+06  4.808e+05   3.964 9.41e-05 ***
factor(AREA)Velachery   -1.056e+05  3.900e+05  -0.271   0.7868
INT_SQFT                 7.926e+02  3.407e+02   2.326   0.0207 *
REG_FEE                  2.088e+01  8.951e-01  23.331  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1259000 on 272 degrees of freedom
Multiple R-squared:  0.8535,    Adjusted R-squared:  0.8492
F-statistic: 198.1 on 8 and 272 DF,  p-value: < 2.2e-16
```

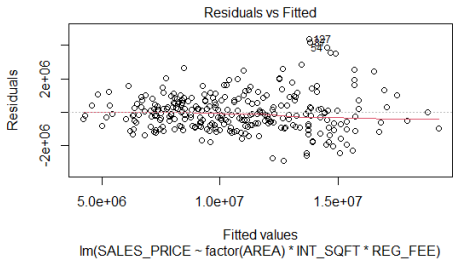
Didapat nilai R squared yang lebih kecil dibandingkan dengan model pertama, yaitu sebesar 0,8535 yang berarti sekitar 85% harga rumah dapat diprediksi dengan luas rumah, biaya registrasi rumah dan lokasi rumah. Bukan berarti model kedua ini lebih kurang merepresentasikan prediksi harga rumah yang digunakan karena hal ini juga dapat terjadi ketika parameter yang digunakan pada model ini lebih sedikit dibandingkan model pertama. Jika kita melihat nilai p-value dari tiap-tiap variabel, dapat dilihat bahwa semuanya berpengaruh secara signifikan kecuali untuk rumah yang berlokasi di Velachery, namun ini dapat terjadi karena Area merupakan variabel kategorik.

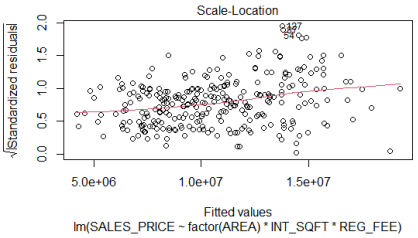
VIF Model:

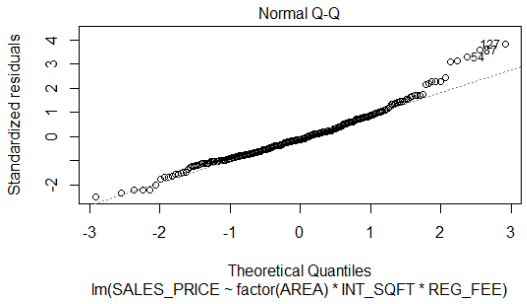
```
                  GVIF Df GVIF^(1/(2*Df))
factor(AREA) 3.989307  6        1.122212
INT_SQFT     4.432476  1        2.105345
REG_FEE      2.162195  1        1.470441
```

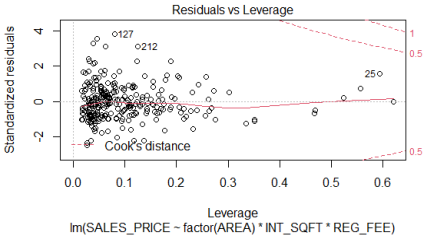
Terlihat pada bagian $GVIF^{(1/(2\times df))}$ nilai yang didapat untuk semua variabel prediktor beserta interaksinya bernilai < 4, yang menandakan model ini telah memenuhi kriteria model terbaik.

Plot **Model**









Terlihat bahwa model memenuhi asumsi normalitas, linearitas, dan homoskedasitas.

Uji Signifikansi Parameter

- Uji T (Parsial)

  Tujuan

  Uji ini dilakukan untuk melihat apakah parameter telah signifikan terhadap variabel yang ada secara individual

  Hipotesis

  $H_0: \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_8 = 0$

  $H_1: \beta_i \neq 0$

  Tingkat Signifikansi

  Tingkat signifikansi yang digunakan pada uji ini adalah 5% = 0.05

  Daerah Penolakan

  Tolak H_0 jika $|t| > t_{\frac{\alpha}{2},(n-p)}$ atau nilai-$p < \alpha$=0.05

  Statistik Uji

```
Call:
lm(formula = SALES_PRICE ~ factor(AREA) + INT_SQFT + REG_FEE,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2750696  -823276   -67225   705203  4232509

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.627e+06  3.938e+05   4.130 4.83e-05 ***
factor(AREA)Anna Nagar   1.565e+06  3.910e+05   4.002 8.11e-05 ***
factor(AREA)Chrompet     1.549e+06  2.669e+05   5.802 1.82e-08 ***
factor(AREA)Karapakam   -1.250e+06  2.850e+05  -4.385 1.66e-05 ***
factor(AREA)KK Nagar    -8.707e+05  4.085e+05  -2.131   0.0340 *
factor(AREA)T Nagar      1.906e+06  4.808e+05   3.964 9.41e-05 ***
factor(AREA)Velachery   -1.056e+05  3.900e+05  -0.271   0.7868
INT_SQFT                 7.926e+02  3.407e+02   2.326   0.0207 *
REG_FEE                  2.088e+01  8.951e-01  23.331  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1259000 on 272 degrees of freedom
Multiple R-squared:  0.8535,    Adjusted R-squared:  0.8492
F-statistic: 198.1 on 8 and 272 DF,  p-value: < 2.2e-16
```

Keputusan

- Dengan p-value $4,83 \times 10^{-5}$, H_0: $\beta_0 = 0$ ditolak
- Dengan p-value $0,0207$, H_0: $\beta_1 = 0$ ditolak
- Dengan p-value $< 2 \times 10^{-16}$, H_0: $\beta_2 = 0$ ditolak
- Dengan p-value $8,11 \times 10^{-5}$, H_0: $\beta_3 = 0$ ditolak
- Dengan p-value $1,82 \times 10^{-8}$, H_0: $\beta_4 = 0$ ditolak
- Dengan p-value $1,66 \times 10^{-5}$, H_0: $\beta_5 = 0$ ditolak
- Dengan p-value $0,0340$, H_0: $\beta_6 = 0$ ditolak
- Dengan p-value $9,41 \times 10^{-5}$, H_0: $\beta_7 = 0$ ditolak
- Dengan p-value $0,7868$, H_0: $\beta_8 = 0$ tidak ditolak

Kesimpulan

Dalam signifikansi 5%, analisis ini mengindikasikan bahwa variabel luas rumah, biaya registrasi rumah, dan area rumah selain yang berlokasi di Velachery berpengaruh secara signifikan terhadap model.

Bab 4
Pengolahan Data dan Analisis Hasil

4.1. Output Final Model

Setelah melalui tahapan seleksi model pada chapter sebelumnya, kami memilih model berikut:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{3A} + \ldots + \beta_8 X_{3F} + \varepsilon \sim NIID$$

sebagai *final model* dimana :

Y : Harga Rumah (Sales Price)
X_1 : Luas Rumah (INT SQFT)
X_2 : Biaya Registrasi Rumah (REG FEE)
X_{3A} : Rumah yang berlokasi di Anna Nagar
X_{3B} : Rumah yang berlokasi di Chrompet
X_{3C} : Rumah yang berlokasi di Karakapam
X_{3D} : Rumah yang berlokasi di KK Nagar
X_{3E} : Rumah yang berlokasi di T Nagar
X_{3F} : Rumah yang berlokasi di Velachery
ϵ : error yang bersifat $NIID(0, \sigma^2)$

berdasarkan alasan-alasan berikut:

1. Berdasarkan prinsip *parsimony* model tersebut termasuk sederhana
2. Nilai R^2 yang sangat baik yaitu mendekati 1. $R^2 = 0,8535$
3. Nilai VIF untuk semua variabel prediktor beserta interaksinya < 10.
4. Model tersebut memenuhi asumsi normalitas, *homoscedasticity*, linearitas, dan model tersebut tidak memiliki multikolinearitas.

4.1.1 Lampiran Output

Dengan menggunakan software R, didapatkan summary dari model, interval kepercayaan 95% untuk koefisien regresi dari masing-masing prediktor sebagai berikut:

```
Call:
lm(formula = SALES_PRICE ~ factor(AREA) + INT_SQFT + REG_FEE,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2750696  -823276   -67225   705203  4232509

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.627e+06  3.938e+05   4.130 4.83e-05 ***
factor(AREA)Anna Nagar  1.565e+06  3.910e+05   4.002 8.11e-05 ***
factor(AREA)Chrompet    1.549e+06  2.669e+05   5.802 1.82e-08 ***
factor(AREA)Karapakam  -1.250e+06  2.850e+05  -4.385 1.66e-05 ***
factor(AREA)KK Nagar   -8.707e+05  4.085e+05  -2.131   0.0340 *
factor(AREA)T Nagar     1.906e+06  4.808e+05   3.964 9.41e-05 ***
factor(AREA)Velachery  -1.056e+05  3.900e+05  -0.271   0.7868
INT_SQFT                7.926e+02  3.407e+02   2.326   0.0207 *
REG_FEE                 2.088e+01  8.951e-01  23.331  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1259000 on 272 degrees of freedom
Multiple R-squared:  0.8535,    Adjusted R-squared:  0.8492
F-statistic: 198.1 on 8 and 272 DF,  p-value: < 2.2e-16
```

Berdasarkan summary dari R, persamaan taksiran regresinya diberikan oleh:

$$\widehat{Y} = 1.627x10^{-6} + 1.565x10^{-6}x_1 + 1.549x10^{-6}x_2 - 1.250x10^{-6}x_{3A} + \ldots + 2.088x10^{-1}x_{3F}$$

4.2 Interpretasi Koefisien Regresi

Final model :

$$\widehat{Y} = 1.627x10^{-6} + 1.565x10^{-6}x_1 + 1.549x10^{-6}x_2 - 1.250x10^{-6}x_{3A} + \ldots + 2.088x10^{-1}x_{3F}$$

Interpretasi :

1. Koefisien Regresi Luas Rumah

   Arti : Ketika nilai luas rumah naik sebesar 1 poin maka akan meningkat harga rumah sebesar $1.565x10^{-6}$, dengan catatan variabel lain tidak berubah atau tetap.

2. Koefisien Regresi Biaya Registrasi Rumah

   Arti : Ketika nilai biaya registrasi rumah naik sebesar 1 poin maka akan meningkat harga rumah sebesar$1.549x10^{-6}$, dengan catatan variabel lain tidak berubah atau tetap.

3. Koefisien Regresi Area Rumah Anna Nagar

   Arti : Ketika nilai area rumah Anna Nagar naik sebesar 1 poin maka akan menurun harga rumah sebesar $1.250x10^{-6}$, dengan catatan variabel lain tidak berubah atau tetap.

.

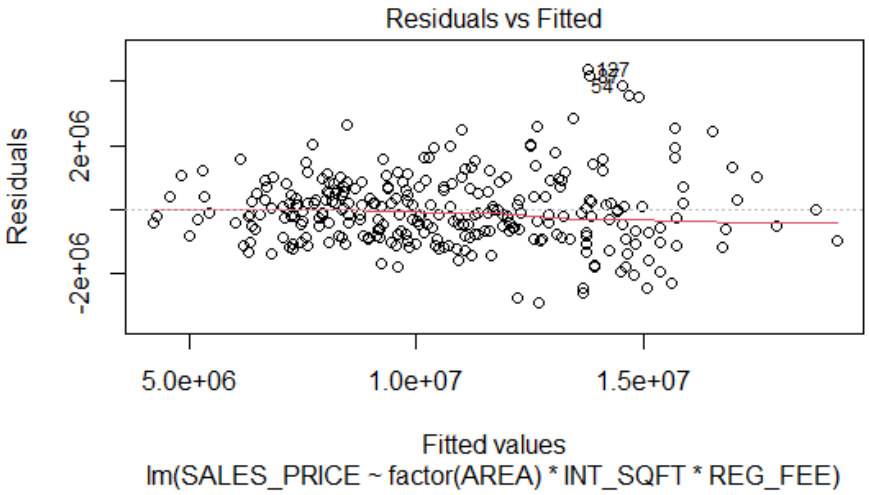
.

. (dst nya hingga)

8. Koefisien Regresi Area Rumah Velachery
   Arti : Ketika nilai area rumah Velachery naik sebesar 1 poin maka akan meningkat harga rumah sebesar $2.088x10^{-1}$, dengan catatan variabel lain tidak berubah atau tetap.

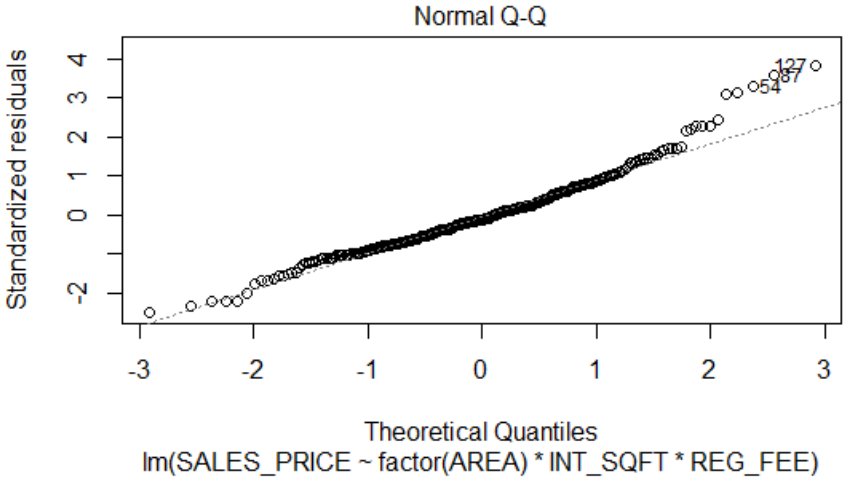
4.3 Analisis Residual

1. Linearitas



Terlihat bahwa titik-titik fitted values terhadap residuals terletak sejalan dengan garis merah yang menandakan model yang kami ajukan memenuhi asumsi linearitas.

2. Normalitas



Berdasarkan Normal Plot di atas, dapat diamati bahwa mayoritas titik-titik data berada di sekitar garis diagonal. Sehingga, dapat disimpulkan bahwa final model yang kami ajukan memenuhi asumsi normalitas. Ada beberapa *outlier* yang memengaruhi Normal Q-Q Plot, seperti data pengamatan ke-54, 87, dan 127.

Cara lain : Dengan menggunakan uji **Kolmogorov-Smirnov**

Hipotesis :

H0 : Residual terdistribusi secara normal

H1 : Residual tidak terdistribusi secara normal

Taraf Signifikansi : $\alpha = 0.05$

Statistik Uji :

```
> ks.test(rstandard(model2), "pnorm")

        One-sample Kolmogorov-Smirnov test

data:  rstandard(model2)
D = 0.07121, p-value = 0.1157
alternative hypothesis: two-sided
```

Aturan Keputusan :

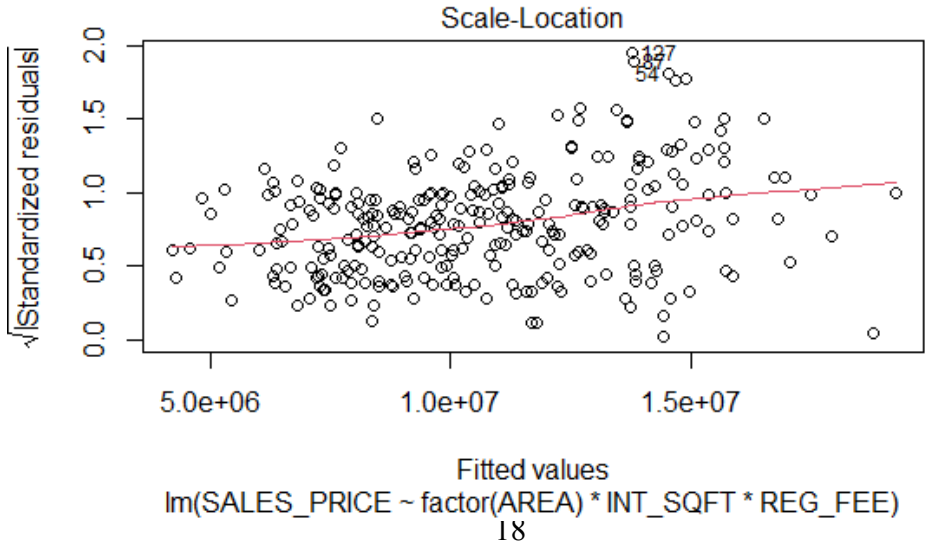
H0 ditolak jika p-value < 0.05

Karena p-value = 0.1157 > 0.05, maka H0 tidak ditolak.

Kesimpulan :

Dengan taraf signifikansi $\alpha = 0.05$, dapat dibuktikan bahwa residualnya berdistribusi normal.

3. Sebaran residual (Homoscedasticity)



Sebaran residualnya tidak membentuk pola tertentu (menyebar secara merata) yang artinya model memenuhi asumsi homoskedatisitas.

4. Nonmultikolinearitas

```
Call:
lm(formula = SALES_PRICE ~ factor(AREA) + INT_SQFT + REG_FEE,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2750696  -823276   -67225   705203  4232509

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.627e+06  3.938e+05   4.130 4.83e-05 ***
factor(AREA)Anna Nagar   1.565e+06  3.910e+05   4.002 8.11e-05 ***
factor(AREA)Chrompet     1.549e+06  2.669e+05   5.802 1.82e-08 ***
factor(AREA)Karapakam   -1.250e+06  2.850e+05  -4.385 1.66e-05 ***
factor(AREA)KK Nagar    -8.707e+05  4.085e+05  -2.131   0.0340 *
factor(AREA)T Nagar      1.906e+06  4.808e+05   3.964 9.41e-05 ***
factor(AREA)Velachery   -1.056e+05  3.900e+05  -0.271   0.7868
INT_SQFT                 7.926e+02  3.407e+02   2.326   0.0207 *
REG_FEE                  2.088e+01  8.951e-01  23.331  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1259000 on 272 degrees of freedom
Multiple R-squared:  0.8535,    Adjusted R-squared:  0.8492
F-statistic: 198.1 on 8 and 272 DF,  p-value: < 2.2e-16
```

Terlihat bahwa standard error dari model semuanya bernilai <1 dan sangat kecil. Hal ini menunjukkan bahwa model tidak memiliki multikolinearitas.

Cara lain :

Salah satu alat statistik untuk menilai multikolinearitas adalah *Variance Inflation Factor* (VIF). Sederhananya, VIF adalah cara untuk mengukur efek multikolinieritas di antara prediktor dalam model kami. VIF < 10 berarti tidak terjadi multikolinearitas antar prediktor.

```
> vif(model2)
                 GVIF Df GVIF^(1/(2*Df))
factor(AREA) 3.989307  6        1.122212
INT_SQFT     4.432476  1        2.105345
REG_FEE      2.162195  1        1.470441
```

Dapat dilihat bahwa VIF semua prediktor < 10. Maka terbukti bahwa model tidak memiliki multikolonieritas.

Berdasarkan analisis residual tersebut, dapat dibuktikan bahwa model yang kami ajukan memenuhi syarat persamaan regresi linie

Bab 5
Penutup

5.1. Kesimpulan

Berdasarkan analisis yang telah dipaparkan sebelumnya, dapat disimpulkan bahwa terdapat hubungan linear antara harga rumah sebagai variabel dependen terhadap biaya registrasi, luas rumah, dan lokasi rumah sebagai variabel prediktor yang digambarkan melalui persamaan taksiran regresi berikut:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{3A} + \ldots + \beta_8 X_{3F} + \varepsilon \sim NIID$$

sebagai *final model* dimana :

Y : Harga Rumah (Sales Price)
X_1 : Luas Rumah (INT SQFT)
X_2 : Biaya Registrasi Rumah (REG FEE)
X_{3A} : Rumah yang berlokasi di Anna Nagar
X_{3B} : Rumah yang berlokasi di Chrompet
X_{3C} : Rumah yang berlokasi di Karakapam
X_{3D} : Rumah yang berlokasi di KK Nagar
X_{3E} : Rumah yang berlokasi di T Nagar
X_{3F} : Rumah yang berlokasi di Velachery
ϵ : error yang bersifat $NIID(0, \sigma^2)$

5.2. Saran

Berdasarkan analisa terhadap hubungan terhadap panjang harga rumah sebagai variabel dependen terhadap biaya registrasi, luas rumah, dan lokasi rumah sebagai variabel prediktor, diperoleh kesimpulan bahwa biaya registrasi, luas rumah, dan lokasi rumah mempengaruhi harga rumah. Dengan begitu, untuk mendapatkan rumah dengan harga murah, kami sarankan untuk memilih rumah dengan luas terkecil dan biaya registrasi termurah. Selain itu, saran kami untuk mendapatkan rumah dengan harga termurah adalah dengan memilih rumah yang berlokasi di Karapakam.

Bab 6
Lampiran

6.1. Folder Kelompok

Folder kelompok dapat diakses pada tautan berikut:
https://drive.google.com/drive/folders/1MSAmVBhdLr081v-tY8GsuAXcwkWiKdFm?usp=sharing

6.2. Sumber Dataset

Sumber dataset dapat diakses pada tautan berikut:
https://www.kaggle.com/datasets/kunwarakash/chennai-housing-sales-price

6.3. Syntax R

```
str(df)

sapply(df, function(x) which(is.na(x)))

df=na.omit(df)

sapply(df, function(x) which(is.na(x)))

unique(df$AREA)

df[df=='Ana Nagar'] = 'Anna Nagar'
df[df=='Ann Nagar'] = 'Anna Nagar'
df[df=='Karapakkam'] = 'Karapakam'
df[df=='Chrompt'] = 'Chrompet'
df[df=='Chrmpet'] = 'Chrompet'
df[df=='Chormpet'] = 'Chrompet'
df[df=='KKNagar'] = 'KK Nagar'
df[df=='TNagar'] = 'T Nagar'
df[df=='Adyr'] = 'Adyar'
df[df=='Velchery'] = 'Velachery'

unique(df$BUILDTYPE)
df[df=='Comercial'] = 'Commercial'
df[df=='Other'] = 'Others'

unique(df$UTILITY_AVAIL)
df[df=='AllPub'] = 'All Pub'
df[df=='NoSewr'] = 'NoSeWa'

unique(df$SALE_COND)
df[df=='Ab Normal'] = 'AbNormal'
df[df=='Partiall'] = 'Partial'
df[df=='PartiaLl'] = 'Partial'
df[df=='Adj Land'] = 'AdjLand'

unique(df$PARK_FACIL)
df[df=='Noo'] = 'No'

unique(df$STREET)
df[df=='Pavd'] = 'Paved'
df[df=='NoAccess'] = 'No Access'

str(df)
library(dplyr)
df=sample_n(df, 300)
View(df)

df2=select(df, c('AREA', 'INT_SQFT','REG_FEE','SALES_PRICE'))
View(df2)

boxplot(df2$REG_FEE,
        main = 'boxplot data variabel TV')
df2 <- droplevels(df2[-which(df2$REG_FEE>700000),])

boxplot(df2$SALES_PRICE,
        main = 'boxplot data variabel TV')
df2 <- droplevels(df2[-which(df2$SALES_PRICE>19000000),])

write.csv(df2, 'DATASET UAS MOLIN.csv', row.names=FALSE)
```

```
df = DATASET.UAS.MOLIN
#Akan dilihat linearitas dari data
plot(df, col="navy", main="Matrix Scatterplot")
```

```r
#model 1 : Dengan Interaksi
sink("Model1.txt")
model1 = lm(SALES_PRICE ~ factor(AREA)*INT_SQFT*REG_FEE, data = df)

summary(model1)
sink()
#model 2 : Tanpa Interaksi
sink('Model2.txt')
model2 = lm(SALES_PRICE~factor(AREA)+INT_SQFT+REG_FEE , data=df)
summary(model2)
sink()

#melihat VIF
library(regclass)
sink('CekModel1.txt')
plot(model1)
VIF(model1)
sink()

sink('CekModel2.txt')
plot(model2)
VIF(model2)
sink()

#Uji Kolmogorov Smirnov untuk cek normalitas
ks.test(rstandard(model2), "pnorm")
```