Project 1

Tujuan:
Tugas ini bertujuan untuk memahami lebih dalam dan menerapkan model regresi linear (sederhana maupun berganda) untuk menyelesaikan permasalahan nyata.

Cakupan tugas:

1. Pemahaman konstekstual, berupa interpretasi kuantitas yang dipelajari dalam konteks contoh permasalahan nyata
2. Pengolahan data, analisis dan interpretasi hasil

Due date: Hari Sabtu di Pekan perkuliahan ke-10 pukul 23.59 WIB

Anggota kelompok:

No	Nama	NPM	Kontribusi	Tingkat kontribusi
1	Aurelio Naufal Effendy	2106638526	Aktif dalam berdiskusi,membuat preprocessing, menyusun laporan dan membuat powerpoint	100%
2	Muhamad Rakan Akmal	2106635745	Aktif dalam berdiskusi, mencari data untuk pemodelan, membuat pemodelan, menuliskan hasil analisis, dan kesimpulan.	100%
3	Musarrofah Kurnia	2106652543	Aktif dalam berdiskusi, menyusun laporan tertulis dan membuat powerpoint	100%
4	Rifa Nayaka Utami	2106632163	Aktif dalam berdiskusi, membuat preprocessing, menyusun laporan tertulis dan membuat powerpoint	100%
5	Shafiyah Audiva Yasmin	2106706880	Aktif dalam berdiskusi, menyusun laporan tertulis dan membuat powerpoint	100%

Instruksi:
Carilah data dengan permasalahan yang dapat dimodelkan dengan regresi linear. Data memuat 200 pengamatan, dengan pengukuran numerik maupun kategorik. Lakukan pengolahan data (jika diperlukan lakukan pre-processing data terlebih dahulu), dengan prosedur yang tepat, kemudian lakukan analisis dan interpretasi hasilnya. Ikuti langkah – langkah berikut.

Bagian 1. Pendahuluan
1.1 Rumusan masalah
Papan, yaitu tempat tinggal atau rumah, menjadi salah satu kebutuhan pokok manusia untuk tempat tinggal, bertahan diri, dan melindungi dari berbagai bahaya, seperti hewan buas, cuaca tidak menentu, dan serangan lain. Setiap tahunnya, angka harapan hidup dan pertumbuhan penduduk meningkat. Hal itu menuntut pemenuhan kebutuhan primer, seperti rumah juga meningkat. Para investor pun berlomba-lomba untuk menjalani bisnis properti, seperti rumah, sebagai sarana investasi, melihat daya beli masyarakat yang melonjak tinggi. Hal itu menjadi salah satu faktor harga rumah setiap tahunnya juga mengalami kenaikan.

Dalam membeli rumah, pembeli rumah tentu ingin membeli rumah dengan rumah yang paling memberikan nilai keuntungan, dengan harga termurah. Nilai keuntungan tersebut bisa berupa banyaknya minimarket di sekitar rumah dan akses transportasi umum yang

mudah dari rumah. Nilai keuntungan yang berbeda-beda tersebut juga menyebabkan harga rumah berbeda-beda. Tidak hanya itu, harga rumah yang berbeda-beda juga dapat disebabkan dari perbedaan tanggal transaksi dan umur rumah. Dengan demikian, kami tertarik menggunakan teknik regresi linier untuk memprediksi harga rumah dengan menganalisis hubungan variabel-variabel tersebut. Hasil dari prediksi ini diharapkan mampu memberikan informasi harga rumah yang sesuai dengan keadaan yang diharapkan.

1.2 Data

A. Permasalahan

Data yang kami ambil merupakan dataset historis pasar dri penilaian dari *real estates* yang dikumpulkan dari Xindian Dist., New Taipei City, Taiwan. Permasalahan ini merupakan permasalahan regresi. Kami akan memprediksi harga *real estates* dari data yang ada dengan menerapkan model regresi linier.

B. Sumber

Kami menggunakan dataset yang bersumber dari laman kaggle dengan linknya sebagai berikut, https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction?datasetId=88705&sortBy=voteCount

C. Ukuran Data

Data memuat 414 pengamatan dengan jumlah pengukuran (kolom data) adalah 8.

D. Skala/tipe Data dan Arti Pengukuran Data Tersebut

Data yang digunakan merupakan dataset, yang berisi 8 jumlah pengukuran, yaitu

No.	Pengukuran	Tipe	Keterangan	Variabel
1.	Nomor rumah	integer (numerik)	merupakan indeks baris, memuat 0-143	dihilangkan (pada tahap *prepocessing*) karena tidak berguna pada pemodelan
2.	X1 tanggal transaksi	float (numerik)	memuat tanggal transaksi rumah dengan bulan dan tahun (4 digit depan adalah tahun dan 3 digit belakang adalah bulan dengan 1 bulan = 83,33), berkisar 2012.667000 - 2013.583000 (Agustus 2012- Juli 2013)	variabel prediktor
3.	X2 umur rumah	float (numerik)	memuat usia rumah dalam tahun, berkisar 0-44 tahun	variabel prediktor
4.	X3 jarak ke stasiun MRT terdekat	float (numerik)	jarak rumah ke stasiun MRT terdekat dalam meter, berkisar 23-6489 meter	variabel prediktor

5.	X4 banyak toko serba ada	integer (kategorik)	banyaknya toko atau minimarket yang dekat dengan rumah, berkisar 0-10	variabel prediktor
6.	X5 lintang	float (numerik)	letak geografis rumah dalam lintang, berkisar 24-26 derajat lintang	dihilangkan (pada tahap *prepocessing*) karena tidak berguna pada pemodelan
7.	X6 bujur	float (numerik)	letak geografis rumah dalam bujur, berkisar 121-122 derajat bujur	dihilangkan (pada tahap *prepocessing*) karena tidak berguna pada pemodelan
8.	Y Harga rumah dari unit area tersebut	float (numerik)	harga rumah per unit area setelah dibagi 10.000 dalam satuan New Dollar Taiwan per ping, berkisar 7.6-117.5 (dikali 10.000 New Dollar Taiwan/Ping)	variabel respon

Maka, data yang akan digunakan dalam pemodelan memuat 414 pengamatan dengan 5 variabel, yaitu 1 variabel respon kuantitatif (Y) , 3 variabel prediktor kuantitatif (X), dan 1 variabel prediktor kualitatif (X).

[Tuliskan dengan baik masalah apa yang akan dibahas, sumber data (*berikan link sumber data, atau lampirkan di Bagian 6)*, ukuran data, dan jumlah pengukuran (kolom data), skala/tipe data, dan arti/maksud dari pengukuran-pengukuran tersebut (jika diketahui). Tentukan variabel respon dan variable prediktor.]

Bagian 2. Pre-processing (jika ada) dan analisis deskriptif

2.1 Load Data

Kami melakukan *import library* yang akan dibutuhkan dalam proses *pre-processing*. Data yang diambil dalam kaggle kami simpan pada Google Drive. Pemanggilan file di Google Drive melalui Google Colab dapat dilihat pada kodingan di bawah ini. Dataset kami simpan pada variabel **df**.

Import modules yang akan dibutuhkan

```
[1] # import modules
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    %matplotlib inline
```

Load data dari Google Drive

```
[2] from google.colab import drive
    drive.mount('/content/drive')
```

```
    Mounted at /content/drive
```

```
[6] # loading data
    df = pd.read_csv('/content/drive/MyDrive/Molin Project 1 Kel 5/Real estate.csv')
```

Berikut head data dari dataset yang telah di-*load*.

Head Dataset

	No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
0	1	2012.917	32.0	84.87882	10	24.98298	121.54024	37.9
1	2	2012.917	19.5	306.59470	9	24.98034	121.53951	42.2
2	3	2013.583	13.3	561.98450	5	24.98746	121.54391	47.3
3	4	2013.500	13.3	561.98450	5	24.98746	121.54391	54.8
4	5	2012.833	5.0	390.56840	5	24.97937	121.54245	43.1

2.2 Mengecek Missing Values

Selanjutnya, kami mengecek apakah kolom atau baris data tersebut mengandung *missing values* atau tidak. Dengan menggunakan *method* **info()**, dapat dilihat bahwa tidak terdapat *missing values* pada data frame **df** .

```
print("\nKeterangan Dataset")
print(df.info())
```

```
Keterangan Dataset
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414 entries, 0 to 413
Data columns (total 8 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   No                                      414 non-null    int64
 1   X1 transaction date                     414 non-null    float64
 2   X2 house age                            414 non-null    float64
 3   X3 distance to the nearest MRT station  414 non-null    float64
 4   X4 number of convenience stores         414 non-null    int64
 5   X5 latitude                             414 non-null    float64
 6   X6 longitude                            414 non-null    float64
 7   Y house price of unit area              414 non-null    float64
dtypes: float64(6), int64(2)
memory usage: 26.0 KB
None
```

2.3 Data Cleaning

Data Cleaning dilakukan untuk memudahkan pemodelan. Pemodelan yang nantinya akan dilakukan adalah teknik regresi linier untuk memprediksi harga rumah dengan menganalisis hubungan variabel-variabel. Variabel-variabel yang akan dianalisis hubungannya hanyalah variabel yang berguna dalam pemodelan. Variabel yang tidak berguna dalam pemodelan, yaitu ‘No’, ‘X5 latitude’, dan ‘X6 longitude’ dihilangkan dalam data.

```
# Variabel yang tidak berguna dalam pemodelan regresi linear akan didrop yakni variabel Nomor rumah, latitude, longitude (karena merupakan bujur dan lintang lokasi rumahnya)
df = df.drop(['No','X5 latitude','X6 longitude'],axis=1)
df.head()
```

	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	Y house price of unit area
0	2012.917	32.0	84.87882	10	37.9
1	2012.917	19.5	306.59470	9	42.2
2	2013.583	13.3	561.98450	5	47.3
3	2013.500	13.3	561.98450	5	54.8
4	2012.833	5.0	390.56840	5	43.1

2.4 Perubahan Nama Variabel dan Penambahan Variabel

Agar tidak menghabiskan banyak tempat dan mudah dipanggil saat melakukan koding, nama-nama kolom pada data **df** akan diubah. Perubahannya adalah sebagai berikut.

1. "**X1 transaction date**" diubah menjadi "**transaction_date**",
2. "**X2 house age**" diubah menjadi "**house_age**",
3. "X**3 distance to the nearest MRT station**" diubah menjadi "**distance_MRT_station**",
4. "**X4 number of convenience stores**" diubah menjadi "**number_conv_stores**", dan
5. "**Y house price of unit area**" diubah menjadi "**houseprice_unit_area**".

```
[ ] # Supaya lebih mudah, nama variabel akan diganti
    df = df.rename(columns = {"X1 transaction date":"transaction_date",
                              "X2 house age": "house_age",
                              "X3 distance to the nearest MRT station": "distance_MRT_station",
                              "X4 number of convenience stores":"number_conv_stores",
                              "Y house price of unit area":"houseprice_unit_area" })
    df.head()
```

	transaction_date	house_age	distance_MRT_station	number_conv_stores	houseprice_unit_area
0	2012.917	32.0	84.87882	10	37.9
1	2012.917	19.5	306.59470	9	42.2
2	2013.583	13.3	561.98450	5	47.3
3	2013.500	13.3	561.98450	5	54.8
4	2012.833	5.0	390.56840	5	43.1

Selanjutnya data yang ada di-*save* dalam bentuk csv untuk pemodelan pada bagian 3.

```
[ ] df.to_csv('data_realestate.csv', index=False) # mensave data yang akan dimasukkan/digunakan untuk membuat model
```

Untuk mendapatkan insight serta visualisasi data yang lebih baik akan dilanjutkan preprocessing dan eksplorasi data analisis (EDA). Akan diambil info berguna dari kolom '**transaction_date**', seperti tahun dan kuartal. Kolom '**transaction_date**' memiliki 2 bagian, yakni Tahun.Kode Bulan. Setiap bulan sama dengan 83,33 unit tambahan dari bulan sebelumnya. Misalnya, 2013.250 berarti Tahun 2013 dengan bulannya adalah 250/83,33 (yaitu bulan ke-3 atau Maret). Maka, dibuatlah dua variabel baru berdasarkan '**transaction_dat**e', yaitu variabel '**transaction_year**' yang berisi tahun pembelian rumah dan '**transaction_month**' yang berisi bulan pembelian rumah. Berdasarkan variabel-variabel tersebut, dibuat variabel quarter, '**transaction_qtr**', yang akan berguna untuk mencari *insight* nantinya.

```
[ ]  # Selanjutnya untuk mendapatkan visualisasi data yang lebih baik akan dilanjutkan preprocessing

     # Akan diambil info berguna dari kolom "transaction date", seperti Tahun, Kuartal.
     # Kolom tanggal transaksi memiliki 2 bagian, yakni Tahun.Kode Bulan
     # Setiap bulan sama dengan 83,33 unit tambahan dari bulan sebelumnya.
     # Misalnya jika 2013.250 berarti -> Tahun adalah 2013 dan bulannya adalah 250/83,33 (yaitu bulan ke-3)

     # Akan dicari list [tahun,kode_bulan].
     year_mon = df['transaction_date'].apply(lambda x:str(x).split("."))  #typecasting setiap nilai ke string dan menspl
     years = [int(year[0]) for year in year_mon] #tahun
     months = [int(year[1]) for year in year_mon] #bulan

     # Membuat kolom baru untuk tahun dan bulan
     df['transaction_year'] = years
     df['transaction_month'] = months

     #Beberapa nilai dalam daftar bulan misalnya: 500 akan diformat otomatis menjadi 5. Angka nol yang dibelakangnya dihapus, sehingga perlu disesuaikan
     df['transaction_month'] = df['transaction_month'].apply(lambda x:x*100 if x<10 else x)
     df['transaction_month'] = df['transaction_month'].apply(lambda x:x*10 if x<100 else x)

     # Bulanya akan dibagi menjadi 4 kuarter
     # (83.3 x 12 bulan = 999.6)
     # Sehingga akan dibuat fungsi untuk mengonversi kode bulan ke Quarter. Misalnya: (0-250->Januari-Maret->Q1, 251-500->April-Juni->Q2, dst..)
     def quarter_conv(month):
         if month <= 250:
             return "Q1"
         elif month <= 500:
             return "Q2"
         elif month <= 750:
             return "Q3"
         elif month <= 1000:
             return "Q4"

     df['transaction_qtr'] = df['transaction_month'].apply(quarter_conv) #Panggil fungsi tersebut

     qtrs = df['transaction_qtr'] #variabel quarter telah dibuat yang akan berguna untuk mencari insight nanti
```

```
df.head()
```

	transaction_date	house_age	distance_MRT_station	number_conv_stores	houseprice_unit_area	transaction_year	transaction_month	transaction_qtr
0	2012.917	32.0	84.87882	10	37.9	2012	917	Q4
1	2012.917	19.5	306.59470	9	42.2	2012	917	Q4
2	2013.583	13.3	561.98450	5	47.3	2013	583	Q3
3	2013.500	13.3	561.98450	5	54.8	2013	500	Q2
4	2012.833	5.0	390.56840	5	43.1	2012	833	Q4

Selanjutnya, kolom '**transaction_date**' dan '**transaction_month**' dapat dihapus karena sudah diwakilkan oleh '**transaction_year'** dan **'transaction_qtr'** ..

2.5 Label Encoding

Sebelum dilakukan visualisasi, perlu dilakukan *label encoding*. *Label encoding* adalah pengubahan setiap label atau nilai-nilai variabel menjadi *unique number* agar dapat dipahami dalam komputasi. Contohnya, saat mencari nilai korelasi antar kolom, komputasi hanya dapat membaca nilai dalam bentuk numerik (*integer* atau *float*), bukan ketegorik. Maka, tidak semua kolom akan dilakukan *labelling*, hanya kolom yang berisi nilai kategorik. Akan dilakukan pengecekan untuk melihat kolom yang mana sajakah yang perlu dilakukan *label encoding*.

```
[ ]  df.dtypes #lihat tipe data

house_age               float64
distance_MRT_station    float64
number_conv_stores        int64
houseprice_unit_area    float64
transaction_year          int64
transaction_qtr          object
dtype: object
```

Dapat dilihat pada kodingan di atas bahwa hanya kolom "**transaction_qtr**" yang tidak bertipe numerik (tidak bertipe *integer* dan *float*). Kolom "**transaction_qtr**" masih bertipe object sehingga diubah menjadi tipe kategorik, lalu diubah lagi menjadi bentuk numerik. Hal itu dilakukan dengan melakukan *dummy encoding*.

```
[ ] # ubah tipe data object ke kategori
    df['transaction_qtr'] = df['transaction_qtr'].astype('category')
    print(df.dtypes) #mengecek apakah sudah benar termasuk data kategori
    df['transaction_qtr'].value_counts() # melihat berapa banyak data yang berada di Q1,Q2,Q3,Q4 dari variabel transaction_qtr
```

```
house_age               float64
distance_MRT_station    float64
number_conv_stores        int64
houseprice_unit_area    float64
transaction_year          int64
transaction_qtr        category
dtype: object
Q2    134
Q4    115
Q1     85
Q3     80
Name: transaction_qtr, dtype: int64
```

```
[ ] # karena variabel "transaction_qtr" dalam tipe 'kategori',sedangkan model ML hanya menerima int atau float. Jadi, akan diubah menjadi int atau float.
    # Dengan Melakukan dummy encoding akan dibentuk df_preprop
    df_dummy = pd.get_dummies(df)
    df_dummy.tail()
```

	house_age	distance_MRT_station	number_conv_stores	houseprice_unit_area	transaction_year	transaction_qtr_Q1	transaction_qtr_Q2	transaction_qtr_Q3	transaction_qtr_Q4
409	13.7	4082.01500	0	15.4	2013	1	0	0	0
410	5.6	90.45606	9	50.0	2012	0	0	1	0
411	18.8	390.96960	7	40.6	2013	1	0	0	0
412	8.1	104.81010	5	52.5	2013	1	0	0	0
413	6.5	90.45606	9	63.9	2013	0	1	0	0

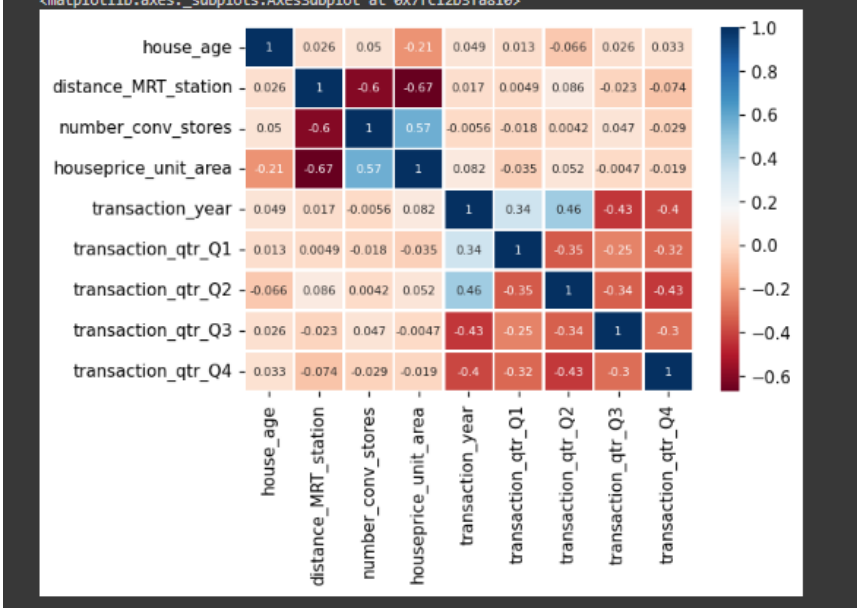
2.6 Visualisasi

1. Heatmap Correlation Matrix

Visualisasi ini bertujuan untuk mempermudah kami melihat korelasi atau hubungan antar variabel yang bersifat numerik atau kuantitatif yang disertai dengan pewarnaan.

```
[ ] #Heatmap setelah di preprocessing
    plt.figure(dpi=110)
    sns.heatmap(df_dummy.corr(), linewidths=1,cmap="RdBu",annot=True, annot_kws={'size':7})
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc12b3fa810>
```



Dapat dilihat bahwa korelasi positif yang sangat kuat ditandai dengan warna biru tua dan korelasi negatif yang sangat kuat ditandai dengan warna merah tua.

Akan dilakukan juga pengurutan korelasi setiap variabel setiap variabel dengan '**houseprice_unit_area**' untuk melihat faktor-faktor apa saja yang paling berkolerasi dengan harga rumah.

```
[ ] # Akan diurutkan korelasi setiap variabel dengan 'houseprice_unit_area' untuk melihat faktor-faktor apa saja yang paling berkolerasi dengan harga rumah
    abs(df_dummy.corr()['houseprice_unit_area']).sort_values()[::-1][:5]

    houseprice_unit_area    1.000000
    distance_MRT_station    0.673613
    number_conv_stores      0.571005
    house_age               0.210567
    transaction_year        0.081545
    Name: houseprice_unit_area, dtype: float64
```

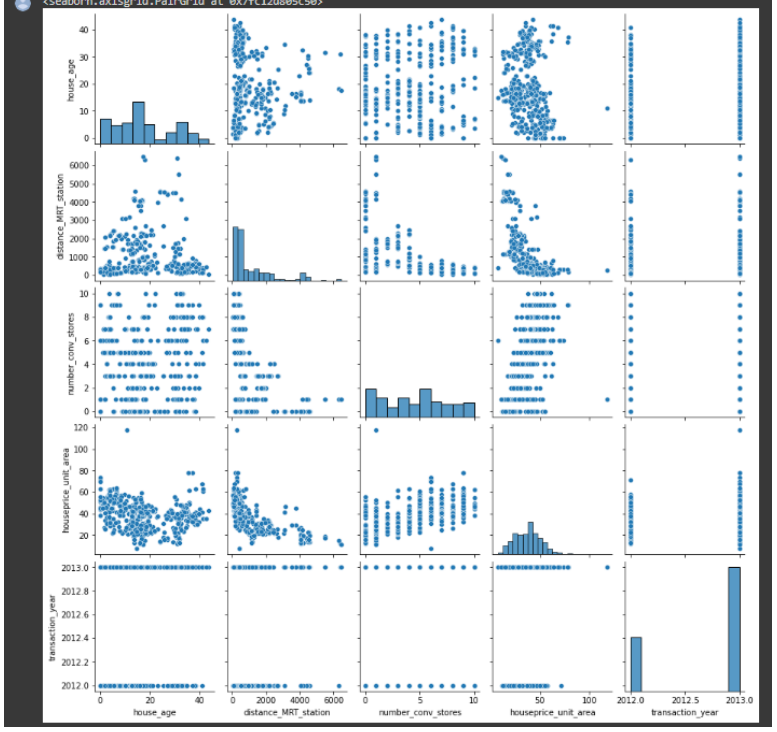
Dari visualisasi dan kodingan tersebut, didapat informasi:

- Terdapat korelasi negatif (sebesar 0.67) antara jarak ke stasiun MRT dengan harga rumah, semakin dekat jarak ke stasiun MRT, harganya semakin mahal, begitu pula sebaliknya,
- Terdapat korelasi positif (sebesar 0.57) antara jumlah banyak toko kecil/mini market di sekitar rumah dengan harga rumah. Semakin banyak toko kecil/mini market di sekitar rumah, maka harga rumahnya semakin mahal, begitu pula sebaliknya.
- Terdapat korelasi negatif (sebesar 0.6) antara jumlah banyak toko kecil/mini market di sekitar rumah dengan jarak ke stasiun MRT, atau dengan kata lain semakin dekat jarak ke stasiun MRT, semakin banyak jumlah toko kecil/mini market.
- Sedangkan umur rumah, dan tahun pembelian rumah korelasinya terhadap harga rumah kecil, sehingga tidak terlalu berpengaruh pada harga rumah.Variabel tahun transaksi sendiri hampir tidak memiliki korelasi sama sekali karena pada data, tahun transaksi hanyalah tahun 2012 dan 2013 saja.
- Orang-orang cenderung untuk membeli rumah pada quarter ke-2, daripada di quarter lain dalam 1 tahun.
- Korelasi tiap-tiap variabel lain terhadap harga rumah tidaklah besar, sehingga model yang akan difitted nantinya pun tidak akan memiliki r-square yang baik.

2. Pairplot

```
# Visualisasi pairplot
sns.pairplot(df)
```

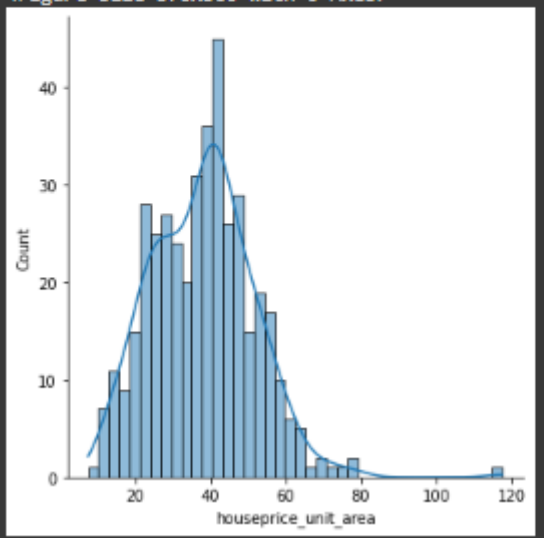
```
<seaborn.axisgrid.PairGrid at 0x7fc12d805c50>
```



Dari pairplot di atas, dapat dilihat distribusi atribut tunggal dan hubungan dua atribut data. Berikut plot distribusi '**houseprice_unit_area**' dalam bentuk yang lebih jelas.

```
[ ]  # Visualisis plot distribusi dari house price of unit area
     plt.figure(figsize=(8,5))
     sns.displot(df['houseprice_unit_area'], bins=40, kde=True)

     <seaborn.axisgrid.FacetGrid at 0x7fc12b5addd0>
     <Figure size 576x360 with 0 Axes>
```



Dapat dilihat bahwa distribusi atau sebaran harga rumah dari unit area tersebut yang membentuk distribusi normal.

Proses *pre-processing* dan visualisasi dalam Google Colab dapat dilihat pada link berikut: https://colab.research.google.com/drive/1g2PFeOJQV1aTs5kQOkQrE-C04igtCSgl?usp=share_link&authuser=4#scrollTo=prtlBKu_D8iu

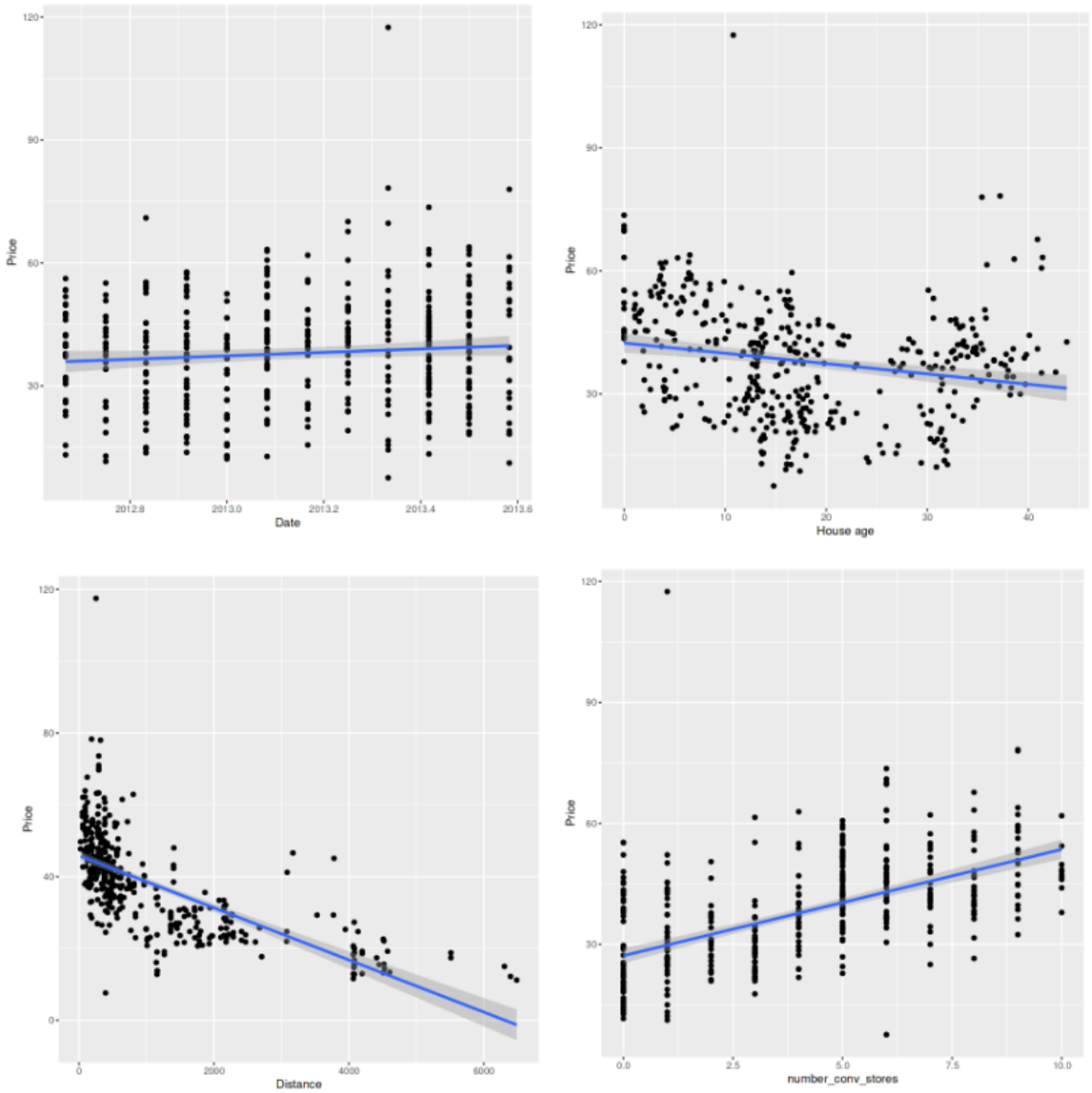
Dari hasil *pre-processing*, kami akan mengajukan 3 model linier yang dapat digunakan untuk melakukan prediksi terhadap '**house_price_unit area' (**keterangan tentang model ada di bagian 3**).** Berikut adalah hipotesis untuk teknik regresi pada proses selanjutnya.

- Hipotesis Model 1
  H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (model tidak berguna)
  H_1: Minimal salah satu dari $\beta_j \neq 0; \; j = 1, 2, 3, 4$ (model berguna)
- Hipotesis Model 2
  H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (model tidak berguna)
  H_1: Minimal salah satu dari $\beta_j \neq 0; \; j = 1, 2, 3, 4, 5$ (model berguna)
- Hipotesis Model 3
  H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ (model tidak berguna)
  H_1: Minimal salah satu dari $\beta_j \neq 0; \; j = 1, 2, 3, 4, 5, 6$ (model berguna)

Bagian 3. Pemodelan

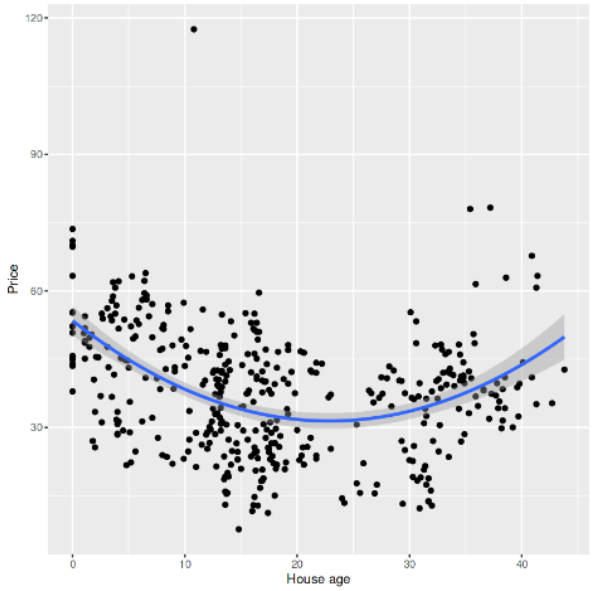
3.1 Plot variabel

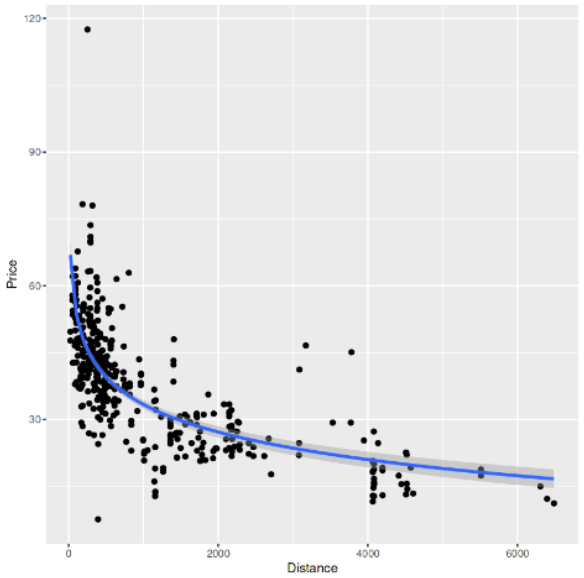
Untuk membuat model yang dapat berjalan dengan baik, maka kita akan melakukan plot setiap data yang menjadi pengaruh nilai y dengan membuat model linier standarnya terlebih dahulu. Apabila data nantinya kurang fitted terhadap model, maka akan dilakukan beberapa cara untuk memperbaharui model.



1. Dari plot untuk transaction_date model dapat dikatakan sudah cukup fitted terhadap data, maka pemodelan ini yang akan digunakan pada model nantinya.
2. Dari plot house_age terlihat bahwa data tersebar hampir menyerupai bentuk kuadratik yang membuat model awal linier kurang fitted terhadap data.
3. Dari plot distance_mrt terlihat bahwa model masih kurang fitted pada data, khusunya pada house age dibawah 2000. Data tersebar menyerupai bentuk logaritma.
4. Dari plot untuk number_conv_stores model dapat dikatakan sudah cukup fitted terhadap data, maka pemodelan ini yang akan digunakan pada model nantinya.

Karena plot (2) dan (3) terlihat masih kurang fitted terhadap data maka akan dilakukan cek model yang lebih mendekati sebaran data untuk (2) dan (3).





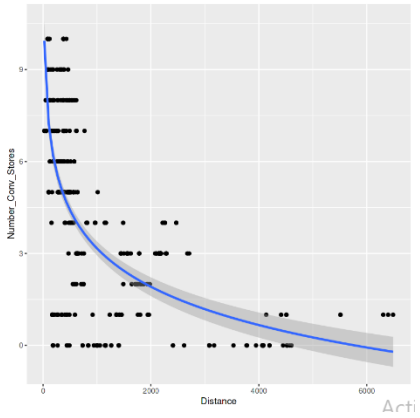
Dapat dilihat sekarang pada plot (2) regresi quadratic ini terlihat data telah cukup fitted terhadap model dan pada plot (3) regresi dengan logaritma ini terlihat data telah cukup fitted terhadap model. Maka untuk pemodelan akan dilakukan dengan model yang menyesuaikan sebaran data pada keempat plot ini.

3.2 Interaksi variabel

Dari plot data yang disajikan di awal, kami mencurigai bahwa adanya interaksi dari num_conv_stores dan distance_MRT_Station. Untuk itu kami akan menelusuri apakah ada interaksi tersebut dengan melakukan Pearson test correlation dan juga melihat plot dari keduanya.

```
        Pearson's product-moment correlation

data:  reg_data$distance_MRT_station and reg_data$number_conv_store
s
t = -15.324, df = 412, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6605398 -0.5373447
sample estimates:
       cor
-0.6025191
```



Dari test pearson untuk korelasi didapatkan nilai p-value < 0.05, maka dapat disimpulkan bahwa korelasi keduanya significant. Terlihat juga pada plot bahwa garis regresi dengan log dari distance_mrt_station dapat dikatakan cukup fitted terhadap sebaran data keduanya. Atas kedua alasan tersebut, kami mempertimbangkan untuk menambahkan adanya interaksi pada model.

3.3 Model yang diajukan

Maka kami mengajukan 3 model linier yang dapat dipakai untuk melakukan prediksi terhadap house_price_unit area.

- Model 1 merupakan model linier standard dengan setiap variabel memiliki nilai coeficientnya masing masing.

- Model 2 kami menambahkan quadratic term untuk house_age dan penggunaan log dari variabel distance_MRT. Hal ini kami lakukan setelah melakukan cek sebaran data dari setiap variabel dan menemukan model yang cocok untuk kedua variabel yang diubah tersebut terhadap persebaran datanya.
- Model 3 hampir sama seperti model 2, tetapi kami menambahkan interaksi dari variabel distance_mrt_station dan num_conv_stores, karena telah dilakukan cek bahwa ada korelasi antara kedua variabel tersebut.

Model 1

$$y = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_4$$

```
Call:
lm(formula = houseprice_unit_area ~ transaction_date + house_age +
    distance_MRT_station + number_conv_stores, data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max
-38.389  -5.630  -0.987   4.306  76.006

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.159e+04  3.215e+03  -3.605 0.000351 ***
transaction_date     5.778e+00  1.597e+00   3.618 0.000334 ***
house_age           -2.545e-01  3.953e-02  -6.438 3.40e-10 ***
distance_MRT_station -5.513e-03  4.480e-04 -12.305  < 2e-16 ***
number_conv_stores   1.258e+00  1.918e-01   6.558 1.65e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.118 on 409 degrees of freedom
Multiple R-squared:  0.5553,    Adjusted R-squared:  0.5509
F-statistic: 127.7 on 4 and 409 DF,  p-value: < 2.2e-16
```

Model 2

$$y = B_0 + B_1x_1 + B_2x_2 + B_3x_{2*} + B_4x_{4*} + B_5x_5$$

$$x_{2*} = (x_2)^2 \; ; \; x_{4*} = \log(x_4)$$

```
Call:
lm(formula = houseprice_unit_area ~ transaction_date + house_age +
    house_age2 + distance_MRT_stationl + number_conv_stores,
    data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max
-34.949  -4.915  -0.786   3.626  73.403

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.504e+04  3.001e+03  -5.012 8.03e-07 ***
transaction_date      7.515e+00  1.491e+00   5.041 6.99e-07 ***
house_age            -8.273e-01  1.466e-01  -5.644 3.11e-08 ***
house_age2            1.536e-02  3.536e-03   4.344 1.77e-05 ***
distance_MRT_stationl -7.078e+00  5.474e-01 -12.932  < 2e-16 ***
number_conv_stores    6.643e-01  1.968e-01   3.375 0.000809 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.451 on 408 degrees of freedom
Multiple R-squared:  0.6189,    Adjusted R-squared:  0.6142
F-statistic: 132.5 on 5 and 408 DF,  p-value: < 2.2e-16
```

Model 3

$$y = B_0 + B_1x_1 + B_2x_2 + B_3x_{2*} + B_4x_{4*} + B_5x_5 + B_6x_4x_5$$

$$x_{2*} = (x_2)^2 \; ; \; x_{4*} = \log(x_4)$$

```
Call:
lm(formula = houseprice_unit_area ~ transaction_date + house_age +
    house_age2 + distance_MRT_stationl + number_conv_stores +
    distance_MRT_stationl * number_conv_stores, data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.642  -4.698  -0.445   3.382  72.228

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -1.485e+04  2.976e+03  -4.989 9.03e-07
transaction_date                        7.423e+00  1.479e+00   5.020 7.74e-07
house_age                              -8.226e-01  1.453e-01  -5.660 2.86e-08
house_age2                              1.519e-02  3.507e-03   4.333 1.86e-05
distance_MRT_stationl                  -8.312e+00  6.966e-01 -11.933  < 2e-16
number_conv_stores                     -1.696e+00  8.578e-01  -1.977  0.04873
distance_MRT_stationl:number_conv_stores 3.913e-01  1.385e-01   2.825  0.00495

(Intercept)                            ***
transaction_date                       ***
house_age                              ***
house_age2                             ***
distance_MRT_stationl                  ***
number_conv_stores                     *
distance_MRT_stationl:number_conv_stores **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.38 on 407 degrees of freedom
Multiple R-squared:  0.6262,    Adjusted R-squared:  0.6207
F-statistic: 113.6 on 6 and 407 DF,  p-value: < 2.2e-16
```

Selanjutnya akan dilakukan cek anova dengan hasil sebagai berikut.

```
anova(reg1, reg2)
anova(reg1, reg3)
anova(reg2, reg3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	409	34002.58	NA	NA	NA	NA
2	407	28580.04	2	5422.538	38.61039	4.427165e-16

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	409	34002.58	NA	NA	NA	NA
2	408	29140.63	1	4861.951	68.07252	2.197454e-15

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	408	29140.63	NA	NA	NA	NA
2	407	28580.04	1	560.5873	7.98316	0.004953918

Telah terbukti bahwa reg1, reg2, dan reg3 merupakan model yang berbeda, karena setelah dilakukan cek anova dengan p-valuenya < 0.05 dan memiliki nilai yang sangat kecil.

3.4 Alasan memilih model

Karena sebelumnya kami telah melihat bahwa adanya interaksi pada model 3 dapat meningkatkan nilai R-squared menjadi yang paling tinggi dari ketiga model. Atas alasan tersebut, kami memilih model 3 untuk dilakukan analisa lebih lanjut.

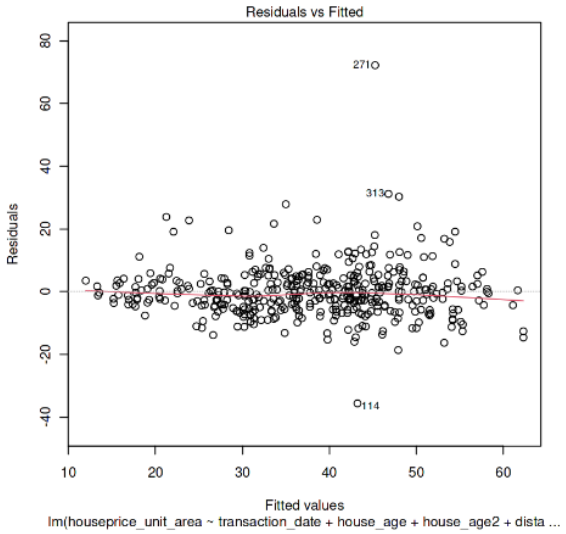
Berikut ini adalah asumsi dari model regresi:

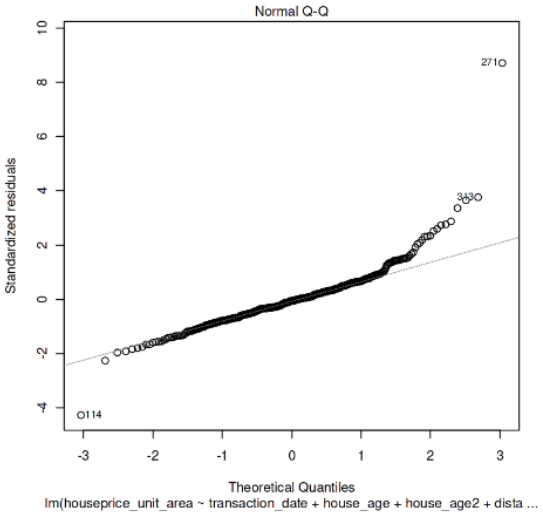
1. Ekspektasi error untuk populasi bernilai 0.
2. Error berdistribusi N(0,1).
3. Variansi error bernilai konstan

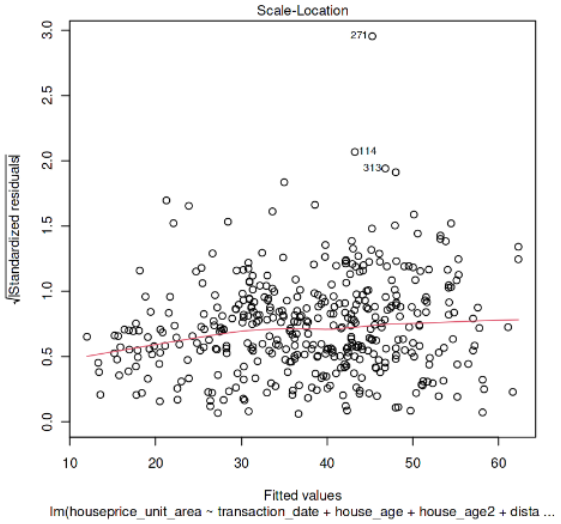
Bagian 4. Pengolahan data dan analisis hasil

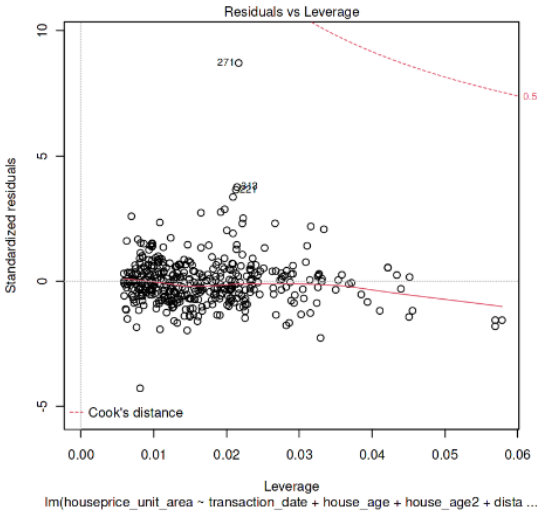
4.1 Asumsi dari model regresi

Dalam analisis model, pertama akan dilakukan cek asumsi dari model regresi:









Didapat bahwa asumsi:

1. Ekspektasi error untuk populasi bernilai 0.
   Pada grafik (1) terlihat bahwa residual dari setiap nilai dugaan model berada di nilai 0, yang berarti asumsi ekspektasi error bernilai 0 untuk populasi dapat diterima.

2. Error berdistribusi N(0,1).
   Pada grafik (2) terlihat bahwa nilai residual standar mendekati nilai standar dari N(0,1) yang digambarkan sebagai garis putus-putus, yang berarti asumsi error berdistribusi N(0,1) dapat diterima.

3. Variansi error bernilai konstan (Homoskedasitas)
   Pada grafik (3) terlihat bahwa akar dari nilai residual standar secara rata-rata tersebar di antara nilai 0 dan 1, yang berarti asumsi bahwa variansi error bernilai konstan dapat diterima.

4.2 Multicollinearity

Pada sebuah model yang memiliki interaksi dan quadratic term, tentunya multicollinearity akan muncul pada variabel-variabel yang bersangkutan. Selanjutnya akan dicari multicollinearity untuk setiap variabel.

```
vif(reg3)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

transaction_date: 1.0223146555991 **house_age:** 16.1246023368379 **house_age2:** 16.0277356294737 **distance_MRT_stationl:** 3.57768196876906 **number_conv_stores:** 37.5462114950729 **distance_MRT_stationl:number_conv_stores:** 28.3692158812376

Dari hasil yang diperoleh hanya variabel transaction_date yang memiliki nilai vif < 5 dan bebas dari indikasi multicollinearity. Variabel house_age memiliki nilai vif yang tinggi karena quadratic term. Distance_MRT_stationl dan number_conv_stores tentunya memiliki nilai yang sangat tinggi karena adanya interaksi pada kedua variabel tersebut.

Multicollinearity yang muncul pada model ini merupakan hal yang normal dan tidak didapat dihindari. Dalam kehidupan nyata tentunya tidak mungkin apabila suatu variabel independen secara murni, pastinya interaksi tersebut akan muncul secara sengaja atau tidak sengaja. Untuk mengetahui bagaimana variabel-variabel tersebut dapat dikatakan independen, maka akan dilakukan cek apabila tidak digunakan quadratic term dan interaksi.

```
reg_check_vif <- lm(houseprice_unit_area~transaction_date +
         house_age + distance_MRT_stationl +
         number_conv_stores, data = reg_data)
vif(reg_check_vif)
```

transaction_date: 1.02107915173931 **house_age:** 1.02232155423244
distance_MRT_stationl: 1.96370054204305 **number_conv_stores:** 1.94140815325905

Dapat terlihat setelah dilakukan cek untuk variabel house_age juga memiliki nilai vif yang sangat kecil yaitu 1.02. Begitu juga dengan distance_MRT_station dan number_conv_stores yang memiliki nilai sedikit lebih besar karena adanya korelasi pada kedua data tersebut.

Terbukti untuk semua variabel memiliki nilai vif yang sangat kecil pada kisaran 1-2, sehingga sebetulnya tidak ada indikasi multicollinearity yang besar atau dapat dikatakan masing-masing variabel independen, sekalipun distance_MRT_station dan num_conv_stores memiliki korelasi.

4.2 Analisis Model

```
Call:
lm(formula = houseprice_unit_area ~ transaction_date + house_age +
    house_age2 + distance_MRT_stationl + number_conv_stores +
    distance_MRT_stationl * number_conv_stores, data = reg_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.642  -4.698  -0.445   3.382  72.228

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -1.485e+04  2.976e+03  -4.989 9.03e-07
transaction_date                         7.423e+00  1.479e+00   5.020 7.74e-07
house_age                               -8.226e-01  1.453e-01  -5.660 2.86e-08
house_age2                               1.519e-02  3.507e-03   4.333 1.86e-05
distance_MRT_stationl                   -8.312e+00  6.966e-01 -11.933  < 2e-16
number_conv_stores                      -1.696e+00  8.578e-01  -1.977  0.04873
distance_MRT_stationl:number_conv_stores  3.913e-01  1.385e-01   2.825  0.00495

(Intercept)                             ***
transaction_date                        ***
house_age                               ***
house_age2                              ***
distance_MRT_stationl                   ***
number_conv_stores                      *
distance_MRT_stationl:number_conv_stores **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.38 on 407 degrees of freedom
Multiple R-squared:  0.6262,    Adjusted R-squared:  0.6207
F-statistic: 113.6 on 6 and 407 DF,  p-value: < 2.2e-16
```

Dengan model yang telah dibuat terlihat bahwa setiap variabel x memiliki nilai p-value < 0.05, sehingga dapat dikatakan seluruhnya sangat signifikan terhadap variabel responnya. Model yang dibuat memiliki nilai adjusted R-squared di sekitar 0.6207. Walaupun nilai tersebut masih berada di bawah 0.75 yang menjadi batas bahwa adanya korelasi yang kuat, model yang dibuat dapat diterima dan dapat digunakan dengan cukup baik.

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ (model tidak berguna)
H_1: Minimal salah satu dari $\beta_j \neq 0;\ j = 1, 2, 3, 4, 5, 6$ (model berguna)

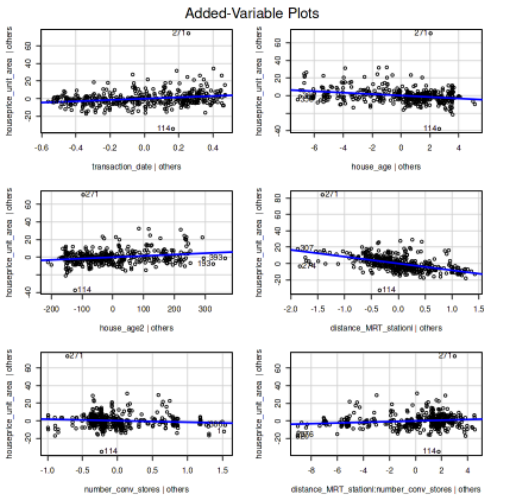
Akan dilakukan uji signifikansi parameter. Dengan nilai p-value setiap variabel yang dibawah 0.05, maka H0 ditolak. Sehingga nilai setiap parameter tidak sama satu dengan yang lainnya dan signifikan terhadap variabel houseprice_unit_area sebagai variabel responnya.

4.2 Korelasi setiap variabel dari model

Selanjutnya kami akan melakukan analisis terhadap korelasi dari masing-masing variabel terlebih dahulu. Dari model yang dibuat, kita dapat mengetahui korelasi dari nilai estimated parameter setiap variabel. Jika estimated parameter bernilai positif, maka hubungannya adalah korelasi positif. Sedangkan jika estimated parameter bernilai negatif, maka hubungannya adalah korelasi negatif.

- Transaction date dengan house price berkorelasi linier positif (Semakin barunya suatu rumah yang dibeli akan meningkatkan nilai house price).
- House age dengan house price berkorelasi linier negatif (Semakin tuanya umur rumah akan menurunkan nilai house price).
- Distance MRT station dengan house price berkorelasi linier negatif (Semakin jauhnya jarak ke MRT Station akan menurunkan nilai house price).
- Number convenience store dengan house price berkorelasi linier negatif (Semakin banyaknya convenience store pada suatu kawasan akan menurunkan nilai house price).
- Interaksi antara Distance MRT station dan Number convenience store dengan house price berkorelasi linier positif (Semakin dekatnya jarak ke MRT station dan semakin banyaknya convenience store pada suatu kawasan akan meningkatkan nilai house price).

```
avPlots(reg3)
```



Untuk analisis korelasi dari house_age, abaikan plot dari house_age2 yang pada gambar terlihat berkorelasi positif. Estimated parameter dari house_age 2 hanya digunakan untuk membantu meningkatkan fitted model terhadap sebaran data. Jika ingin melihat korelasi untuk house_age dapat hanya melihat plot dari house_age yang berkorelasi negatif. Jadi dapat disimpulkan bahwa plot dari estimated parameter ini menguatkan analisis korelasi yang dibuat sebelumnya adalah benar.

4.3 Kejanggalan dalam model

Pada hubungan tersebut terlihat ada 2 kejanggalan:

1. Variabel interaksi antara Distance MRT station dan Number convenience store lebih signifikan daripada Number convenience store.
2. Banyaknya convenience store yang justru menurunkan nilai house price.

Jika kita melihat dalam konteks kehidupan nyata, tentunya di sekitaran MRT Station terdapat banyak convenience store. Pembangunan suatu convenience store akan mengikuti kebutuhan dari bagaimana seseorang beraktivitas setiap hari dan area yang menjadi pusat aktivitas. Hal itu yang menjadi alasan mengapa terdapat banyak convenience store di MRT station atau stasiun transportasi, karena merupakan area yang menjadi pusat aktivitas seseorang setiap harinya. Setelah melihat konteks ini, kejanggalan tersebut dapat terjawab dengan baik, karena adanya suatu MRT station pasti disusul dengan adanya convenience store pada area tersebut.

Variabel convenience store pada model ini memiliki p-value yang hampir mendekati 0.05, sehingga dapat dikatakan tidak signifikan kuat. Seperti yang telah dijelaskan sebelumnya, bagaimana pembangunan convenience store dipengaruhi oleh MRT Station, pada umumnya di beberapa area yang tidak dijangkau MRT station seperti di daerah bukan perkotaan, convenience store merupakan salah satu hal yang paling utama untuk memenuhi kebutuhan sehari-harinya. Karena hal itu, dapat dikatakan bahwa semakin banyak convenience store yang tidak dipengaruhi oleh MRT station, maka area tersebut diasumsikan jauh dari perkotaan. Kejanggalan banyaknya convenience store yang menurunkan nilai house price, menjadi dapat diterima ketika model memiliki interaksi.

4.4 Variabel yang paling mempengaruhi

Dari model yang dibuat, kita dapat memberikan suggesstion terkait variabel apa yang paling mempengaruhi tingginya harga suatu rumah. Untuk mencari best predictor variable, kita tidak dapat semata-mata melihat besarnya estimated parameter/coeficient dari model yang dibuat. Akan dilakukan analisis Relative Importance Method untuk melihat variabel yang paling mempengaruhi pada model.

```
regressor <- lm(houseprice_unit_area~transaction_date +
        house_age + house_age2 + distance_MRT_stationl +
        number_conv_stores + distance_MRT_stationl*number_conv_stores,
          data = reg_data)
relImportance <- calc.relimp(regressor, type = "lmg", rela = TRUE)
sort(relImportance$lmg, decreasing=TRUE)
```

distance_MRT_stationl: 0.578580696394823 **number_conv_stores:** 0.265723453484553 **house_age:** 0.0737579004792338 **house_age2:** 0.0415510473801225 **transaction_date:** 0.0283086490913733 **distance_MRT_stationl:number_conv_stores:** 0.0120782531698936

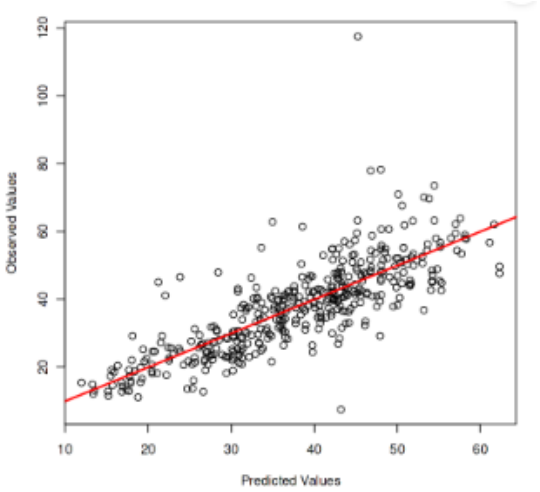
Dari hasil ini dapat terlihat bahwa variabel yang paling mempengaruhi harga rumah pada suatu daerah adalah distance_MRT_Stationl (jarak rumah tersebut dengan stasiun MRT). Kemudian variabel yang paling mempengaruhi selanjutnya adalah number_conv_stores (banyaknya toko perbelanjaan), house_age (umur dari suatu rumah),

dan yang paling terakhir adalah transaction_date (atau waktu ketika transaksi tersebut dilaksanakan), serta interaksi distance_MRT_stationl dengan num_conv_stores.

Dalam variabel distance MRT Station dan number convenience store menjadi data yang paling berpengaruh dikarenakan pada umumnya dalam kebutuhan orang yang memiliki rumah, jarak antara tempat tinggal dengan stasiun untuk melakukan transportasi ke tempat kerja ataupun tempat lainnya, dan banyaknya toko perbelanjaan untuk memenuhi kebutuhan hidupnya adalah hal yang paling utama. Hal itu yang membuat rumah di kota-kota besar memiliki harga yang begitu mahal, karena tentunya pembangunan stasiun transportasi dan toko perbelanjaan akan lebih terpusat di kota-kota besar.

Yang menjadikan transaction date menjadi variabel yang paling tidak berpengaruh pada model, selain dari variabel interaksi adalah karena pada data model ini, transaction date hanya tersebar pada tahun 2012 - 2013. Dalam rentang waktu tersebut, tentunya harga suatu rumah tidak akan mengalami peningkatan yang begitu besar. Tidak seperti dimisalkan perbandingan rumah pada tahun 1950 dengan tahun 2022 yang tentunya akan mengalami peningkatan begitu pesat atas alasan ekonomi dan sebagainya.

Bagian 5. Penutup



Dari plot ini terlihat bahwa nilai observed values dapat dikatakan telah fitted terhadap model atau predicted valuesnya. Dapat disimpulkan bahwa model yang dibuat dengan quadratic terms dan juga menambahkan interaksi dapat menjawab permasalahan dalam melakukan prediksi harga rumah pada suatu kawasan. Dari analisis yang dilakukan, setiap variabel yang membangun model adalah signifikan atau berpengaruh terhadap variabel responnya.

Hasil analisis yang dilakukan juga dapat menjawab pertanyaan variabel yang paling mempengaruhi harga rumah pada suatu daerah, yaitu distance_MRT_Stationl (jarak rumah tersebut dengan stasiun MRT). Kemudian variabel yang paling mempengaruhi selanjutnya adalah number_conv_stores (banyaknya toko perbelanjaan), house_age (umur dari suatu rumah), dan yang paling terakhir adalah transaction_date (atau waktu ketika transaksi tersebut dilaksanakan), serta interaksi distance_MRT_stationl dengan num_conv_stores.

Bagian 6. Lampiran
Link Google Drive :
https://drive.google.com/drive/folders/1cPDlgzjpJ_cpIa2N_AESxWsLzSWdMOK9?usp=share_link

Link Google Colab (Prepocessing) :
https://colab.research.google.com/drive/1g2PFeOJQV1aTs5kQOkQrE-C04igtCSgl?usp=share_link

Link Kaggle (Model dan analisis) :
https://www.kaggle.com/code/muhamadrakanakmal/group-5-model-linier/notebook

Link Youtube presentasi :
https://youtu.be/-SpZGcCFcdg