Judul : ***Estimasi Harga Properti Berdasarkan Luas Tanah, Luas Bangunan, dan Fitur Rumah Bulan November 2024 di Jakarta***

Anggota : **1. Salma Hanifa (123220019) (DS-E)**
2. Diandra Yusuf Arrafi (123220031) (DS-E)
3. Fahmi Kariem (123220028) (DS-E)

1. Business Understanding & Analytic Approach

Deskripsi Masalah

Pasar properti seringkali menghadapi ketidakpastian harga akibat berbagai faktor seperti lokasi, ukuran tanah dan bangunan, jumlah kamar. Hal ini menciptakan tantangan bagi pembeli yang kesulitan menilai harga, penjual yang sulit menentukan harga optimal, dan investor yang memerlukan data akurat untuk investasi yang menguntungkan. Ketidakpastian ini menghambat efisiensi proses jual beli, meningkatkan risiko kesalahan harga, dan memperlambat pengambilan keputusan. Tantangan Utama juga meliputi kurangnya transparansi pasar tanpa standar harga yang jelas, data pasar yang tidak konsisten, serta keputusan yang sering kali didasarkan pada opini, bukan data, sehingga berpotensi mengakibatkan transaksi yang tidak optimal.

Usulan Task

Task yang kami usulkan adalah Prediksi Harga Properti berdasarkan data properti seperti lokasi, luas tanah dan bangunan, jumlah kamar. Task ini dipilih karena dapat mengatasi masalah ketidakpastian harga dengan menggunakan data relevan untuk membangun model prediksi yang sesuai, serta cocok diterapkan dengan supervised learning yang memiliki target harga (price). Solusi yang diusulkan adalah membangun model prediksi berbasis machine learning, seperti Random Forest untuk mempelajari pola data historis dan memberikan estimasi harga properti secara akurat.

Mendefinisikan Business Understanding tentang permasalahan ketidakpastian harga properti yang dihadapi pembeli, penjual, dan investor, serta dampaknya terhadap efisiensi pasar.
Menentukan Analytic Approach yang digunakan. Pada topik ini, kami menggunakan pendekatan analitik berbasis prediksi (Predictive Analysis) untuk memodelkan harga properti.
Menganalisis dan mengumpulkan data yang relevan, seperti luas tanah, luas bangunan, jumlah kamar, lokasi, dan tipe rumah, dengan sumber data dari situs properti terpercaya.
Menentukan jenis model yang digunakan. Pada topik ini, kami menggunakan model Supervised Learning berbasis prediksi untuk memprediksi harga properti berdasarkan pola data.
Menentukan Algoritma yang digunakan. Pada topik ini, kami menggunakan algoritma Random Forest Regressor, yang efektif untuk menangani data non-linear dan campuran numerik serta kategorikal.
Mendeskripsikan hasil yang telah didapat, termasuk evaluasi performa model menggunakan metrik seperti R^2 Score, MAE, dan RMSE untuk menilai akurasi dan kesalahan prediksi, serta memberikan rekomendasi berdasarkan hasil analisis.

Data Science Metodologi:

1. **Data Collection**
   Memastikan data yang terkumpul mencakup informasi lengkap untuk memprediksi harga rumah.
2. **Preprocessing Data**
   - Pembersihan Data: Menghapus data duplikat dan menghapus data yang tidak valid (contoh: sertifikasi = "Tidak Tersedia")
   - Transformasi Data: Mengekstrak angka dari kolom luas tanah, luas bangunan, kamar tidur dan mengubah kolom price menjadi format numerik (hilangkan "Rp", "Jt", "M").
   - Encoding Data Kategorikal: Mengubah kolom seperti type, loc, dan sertifikasi menjadi numerik dengan One-Hot Encoding.
3. **Exploratory Data Analysis (EDA)**
   - Distribusi Harga Rumah.
   - Hubungan antara Fitur dan Harga.
   - Analisis Berdasarkan Kategori.
   - Korelasi Antar Fitur.
4. **Modeling**
   - Split Dataset.
   - Gunakan Random Forest Regressor.
   - Latih Model.
   - Hyperparameter Tuning (Opsional).
5. **Evaluation**
   - Prediksi harga pada test set.
   - Hitung MAE, MSE, dan R^2 Score.
   - Analisis hasil prediksi (Bandingkan harga aktual dan prediksi pada test set).

2. Data Requirement & Data Collection

Data yang dikumpulkan merupakan data penjualan rumah di wilayah Jakarta dari hasil scrapping di website www.pinhome.id pada tanggal 02 Desember 2024 pukul 22:50 WIB. Data yang kita ambil terdiri dari:

- Date (masih dalam bentuk text kapan terakhir diperbaharui)
- Type (Baru / Second)
- Price (berupa text harga)
- Loc (kecamatan & kota, contoh: Cakung, Kota Jakarta Timur)
- kamar_tidur (berupa text jumlah kamar tidur)
- luas_bangunan (berupa text luas bangunan)
- luas_tanah (berupa text luas tanah)
- sertifikasi (SHM, SGB, Strata, Girik)

Preview Data:

- Sebelum dilakukan cleaning data :

date	type	price	loc	kamar_tidur	luas_tanah	luas_bangunan	sertifikasi
Diperbarui 7 menit lalu	Rumah Second	Rp2,9 M	Grogol Petamburan, Kota Jakarta Barat	2 KT	LT 82m²	LB 200m²	SHM
Diperbarui 8 menit lalu	Rumah Second	Rp675 Jt	Pulogadung (Pulo Gadung), Kota Jakarta Timur	2 KT	LT 60m²	LB 60m²	SHM
Diperbarui 15 menit lalu	Rumah Second	Rp2,5 M	Cakung, Kota Jakarta Timur	4 KT	LT 112m²	LB 160m²	HGB
Diperbarui 16 menit lalu	Rumah Second	Rp2,9 M	Grogol Petamburan, Kota Jakarta Barat	3 KT	LT 82m²	LB 200m²	SHM
Diperbarui 18 menit lalu	Rumah Second	Rp22 M	Kebayoran Lama, Kota Jakarta Selatan	4 KT	LT 413m²	LB 700m²	SHM
Diperbarui 20 menit lalu	Rumah Second	Rp1,5 M	Kramatjati (Kramat Jati), Kota Jakarta Timur	2 KT	LT 62m²	LB 84m²	HGB
Diperbarui 20 menit lalu	Rumah Second	Rp2,9 M	Cakung, Kota Jakarta Timur	3 KT	LT 98m²	LB 134m²	SHM
Diperbarui 21 menit lalu	Rumah Second	Rp475 Jt	Jatinegara, Kota Jakarta Timur	3 KT	LT 60m²	LB 60m²	SHM
Diperbarui 23 menit lalu	Rumah Second	Rp5,5 M	Cakung, Kota Jakarta Timur	6 KT	LT 336m²	LB 450m²	SHM

- Setelah dilakukan cleaning dan transformasi data :

date	type	price	kamar_tidur	luas_tanah	luas_bangunan	sertifikasi	kecamatan	kota
2024-12-02	Rumah Second	2900	2	82	200	SHM	Grogol Petamburan	Kota Jakarta Barat
2024-12-02	Rumah Second	675	2	60	60	SHM	Pulogadung (Pulo Gadung)	Kota Jakarta Timur
2024-12-02	Rumah Second	2500	4	112	160	HGB	Cakung	Kota Jakarta Timur
2024-12-02	Rumah Second	2900	3	82	200	SHM	Grogol Petamburan	Kota Jakarta Barat
2024-12-02	Rumah Second	22000	4	413	700	SHM	Kebayoran Lama	Kota Jakarta Selatan
2024-12-02	Rumah Second	1500	2	62	84	HGB	Kramatjati (Kramat Jati)	Kota Jakarta Timur
2024-12-02	Rumah Second	2900	3	98	134	SHM	Cakung	Kota Jakarta Timur
2024-12-02	Rumah Second	475	3	60	60	SHM	Jatinegara	Kota Jakarta Timur
2024-12-02	Rumah Second	5500	6	336	450	SHM	Cakung	Kota Jakarta Timur

Sebaran Data:

- Jumlah Data:
  - Sebelum dilakukan cleaning data : 8997 Data
  - Setelah dilakukan cleaning data : 6816 Data

- Jumlah Kolom:
  - Sebelum dilakukan cleaning data : 8 Kolom
    (date, type, price, kamar_tidur, luas_tanah, luas_bangunan, sertifikasi, loc)
  - Setelah dilakukan cleaning data : 9 Kolom
    (date, type, price, kamar_tidur, luas_tanah, luas_bangunan, sertifikasi, kecamatan, kota)

(penjelasan dari sebelum dan sesudah cleaning data ada pada bagian data preparation).

Penggunaan Data:

Data yang digunakan untuk melakukan prediksi adalah:

- Data Target : price
- Data Parameter : kamar_tidur, luas_tanah, luas_bangunan, kota

Variabel-variabel diatas penting dan merupakan aspek utama yang kami pikirkan dalam menentukan harga rumah. Meski demikian, kami tidak mempertimbangkan sertifikasi sebagai faktor untuk memprediksi harga rumah.

3. Data Preparation

Pada bagian Preparation,hal pertama yang dilakukan adalah melakukan cleaning dan transformasi data. Seperti yang dipaparkan pada Data Requirement & Data Collection, pada proses cleaning dibagi menjadi beberapa tahap, antara lain :

1. Menghapus semua baris yang terdapat data na (null / kosong).
2. Menghapus data yang duplikat.

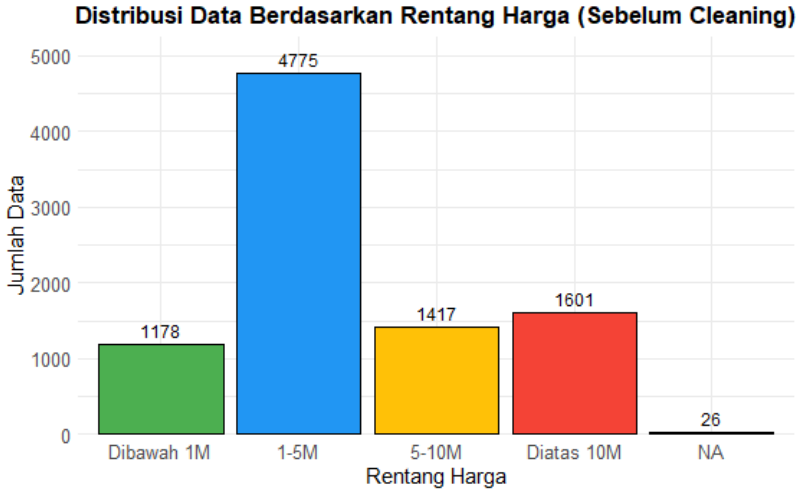
3. Pada data price, menghapus baris yang data pricenya berupa range (misal: 120jt - 5m).
4. Pada data price, mengkonversi data string menjadi int dalam satuan juta (misal: 120jt menjadi 120, 5m menjadi 5000).
5. Pada data date, sebelumnya merupakan data terakhir diperbarui (misal: terakhir diperbarui 2 hari lalu), oleh karena itu dilakukan konversi dari data string menjadi data date berdasarkan tanggal sesuai dengan waktu proses scraping dilakukan yaitu 2 Desember 2024 (misal: terakhir diperbarui 2 hari lalu, maka hasil setelah dikonversi menjadi 30 November 2024).
6. Pada data luas tanah, luas bangunan, jumlah kamar tidur, diubah dari yang tadinya berupa string diubah menjadi angka saja (misal: 2 KT menjadi 2).
7. Pada data loc, yang tadinya kecamatan digabung dengan kota (misal: Cakung, Kota Jakarta Timur), diubah menjadi 2 data, yaitu kecamatan dan kota (misal: data kecamatan = Cakung, data kota = Kota Jakarta Timur). Lalu data loc dihapus.
8. Preview data sebelum dilakukan cleaning (Jumlah data 8997 data) :

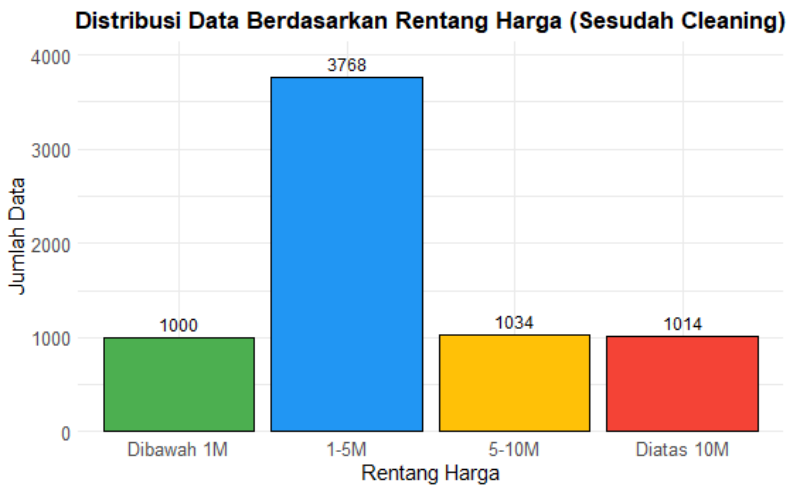
date	type	price	loc	kamar_tidur	luas_tanah	luas_bangunan	sertifikasi
Diperbarui 7 menit lalu	Rumah Second	Rp2,9 M	Grogol Petamburan, Kota Jakarta Barat	2 KT	LT 82m²	LB 200m²	SHM
Diperbarui 8 menit lalu	Rumah Second	Rp675 Jt	Pulogadung (Pulo Gadung), Kota Jakarta Timur	2 KT	LT 60m²	LB 60m²	SHM
Diperbarui 15 menit lalu	Rumah Second	Rp2,5 M	Cakung, Kota Jakarta Timur	4 KT	LT 112m²	LB 160m²	HGB
Diperbarui 16 menit lalu	Rumah Second	Rp2,9 M	Grogol Petamburan, Kota Jakarta Barat	3 KT	LT 82m²	LB 200m²	SHM
Diperbarui 18 menit lalu	Rumah Second	Rp22 M	Kebayoran Lama, Kota Jakarta Selatan	4 KT	LT 413m²	LB 700m²	SHM
Diperbarui 20 menit lalu	Rumah Second	Rp1,5 M	Kramatjati (Kramat Jati), Kota Jakarta Timur	2 KT	LT 62m²	LB 84m²	HGB
Diperbarui 20 menit lalu	Rumah Second	Rp2,9 M	Cakung, Kota Jakarta Timur	3 KT	LT 98m²	LB 134m²	SHM
Diperbarui 21 menit lalu	Rumah Second	Rp475 Jt	Jatinegara, Kota Jakarta Timur	3 KT	LT 60m²	LB 60m²	SHM
Diperbarui 23 menit lalu	Rumah Second	Rp5,5 M	Cakung, Kota Jakarta Timur	6 KT	LT 336m²	LB 450m²	SHM

9. Preview data sesudah dilakukan cleaning (Jumlah data 6816 data) :

date	type	price	kamar_tidur	luas_tanah	luas_bangunan	sertifikasi	kecamatan	kota
2024-12-02	Rumah Second	2900	2	82	200	SHM	Grogol Petamburan	Kota Jakarta Barat
2024-12-02	Rumah Second	675	2	60	60	SHM	Pulogadung (Pulo Gadung)	Kota Jakarta Timur
2024-12-02	Rumah Second	2500	4	112	160	HGB	Cakung	Kota Jakarta Timur
2024-12-02	Rumah Second	2900	3	82	200	SHM	Grogol Petamburan	Kota Jakarta Barat
2024-12-02	Rumah Second	22000	4	413	700	SHM	Kebayoran Lama	Kota Jakarta Selatan
2024-12-02	Rumah Second	1500	2	62	84	HGB	Kramatjati (Kramat Jati)	Kota Jakarta Timur
2024-12-02	Rumah Second	2900	3	98	134	SHM	Cakung	Kota Jakarta Timur
2024-12-02	Rumah Second	475	3	60	60	SHM	Jatinegara	Kota Jakarta Timur
2024-12-02	Rumah Second	5500	6	336	450	SHM	Cakung	Kota Jakarta Timur

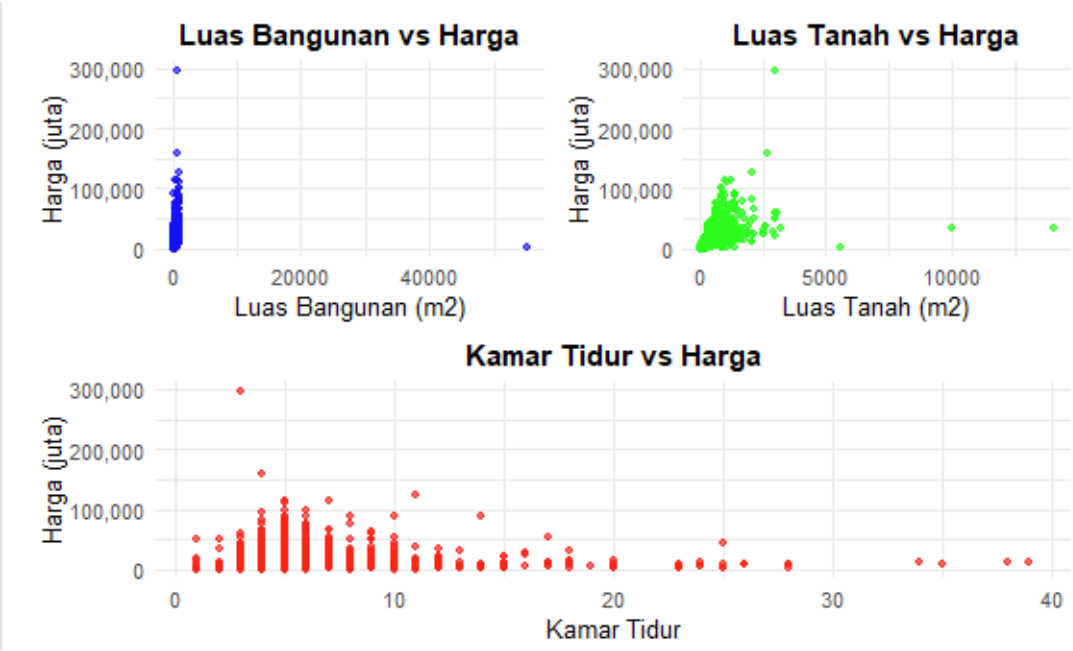
Data cleaning dilakukan untuk menyaring dan mentransformasi data supaya data yang nantinya digunakan untuk modeling dapat lebih jelas. Bisa dilihat perbandingan persebaran data sebelum dan sesudah cleaning berdasarkan range harga (di bawah 1m, 1-5m, 5 - 10m, diatas 10m) :

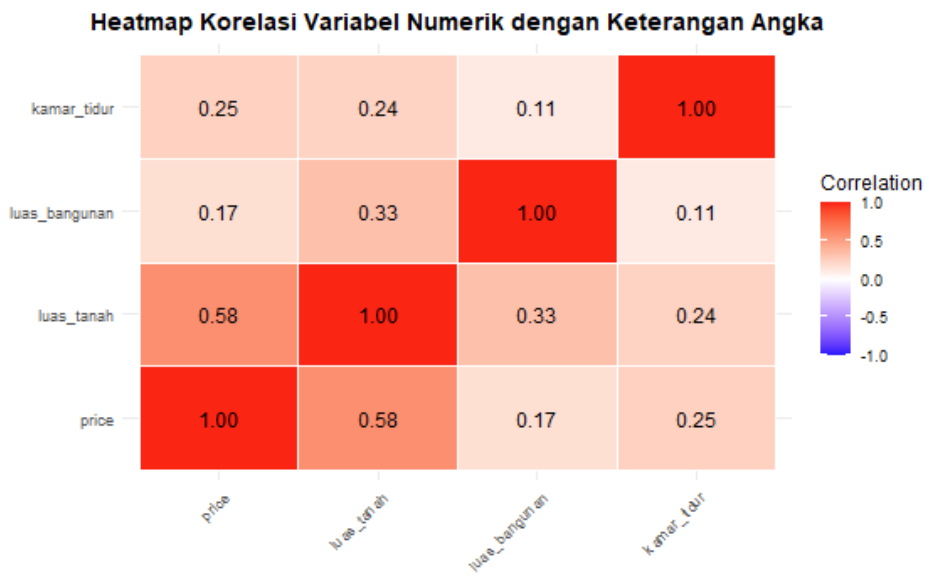




Dari perbandingan data diatas, diketahui bahwa terdapat 26 data NA sebelum dilakukan cleaning (belum termasuk data yang harganya berupa range). Setelah dilakukan cleaning, terlihat bahwa data rumah dengan harga antara 1-5M merupakan harga mayoritas dengan 3768 data, dimana data lainnya kompak berada disekitar 1000an data.

Setelah tahap cleaning data, dilakukan visualisasi Scatter Plot dan Heatmap untuk melihat korelasi antara semua data parameter numerik dengan data harga :

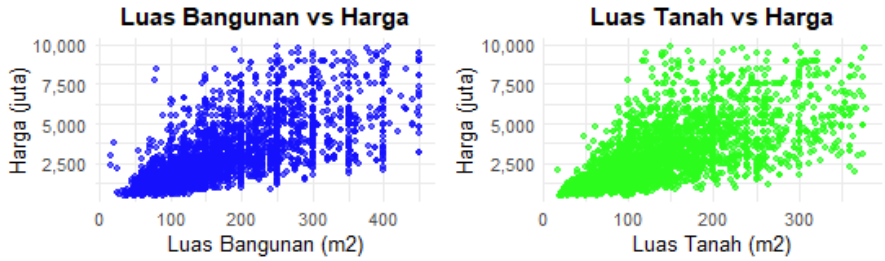


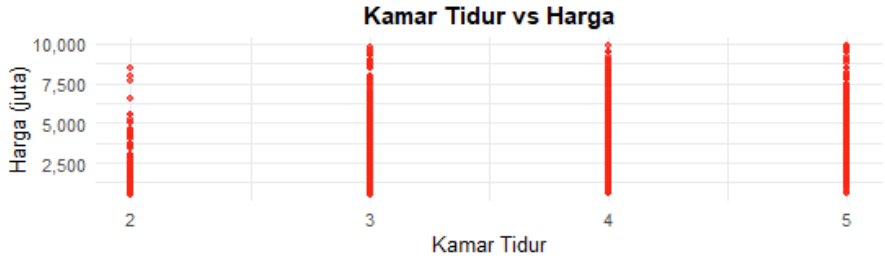


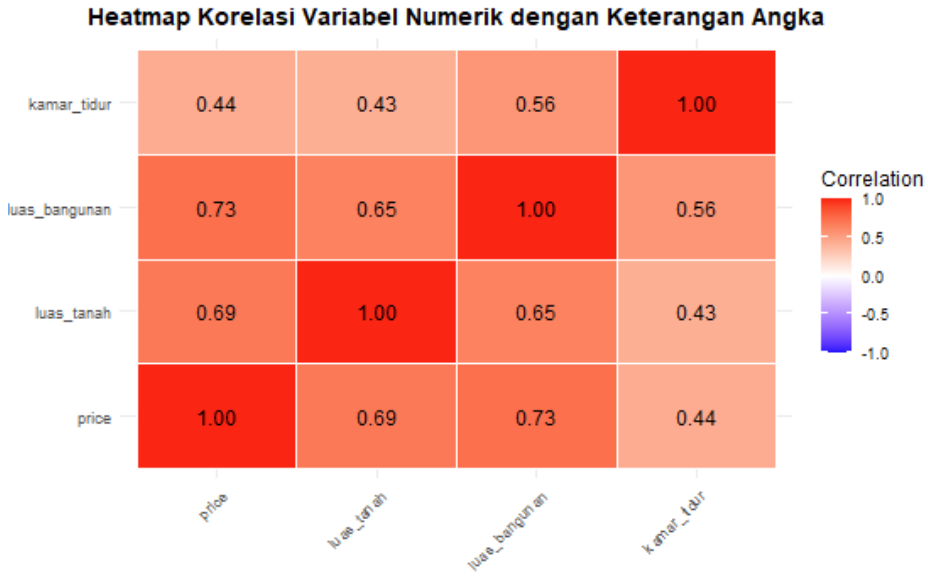
Dari ketiga Scatter Plot tersebut, bisa dilihat terdapat data outlier yang bisa dihapus sebelum masuk ke dalam tahap modelling. Dan juga berdasarkan heatmap, bisa dilihat bahwa sebanyak 58% harga rumah dipengaruhi oleh luas tanah, yang berarti luas tanah memiliki peran yang penting dalam menentukan harga rumah sedangkan data luas bangunan dan kamar tidur memiliki korelasi dibawah 50% dengan harga. Setelah berhasil mendapatkan informasi tersebut, kita mulai melanjutkan proses untuk menghapus outlier.

Pada kasus ini kita menggunakan metode IQR[1] untuk menghapus outlier, metode ini kita pilih karena metode ini merupakan cara yang efektif untuk mengidentifikasi dan menghapus outlier tanpa terpengaruh oleh nilai-nilai ekstrim yang mungkin ada di dataset. Metode ini berfokus pada rentang tengah data, yaitu antara kuartil pertama (Q1) dan kuartil ketiga (Q3), yang mencerminkan distribusi data utama. Dengan menetapkan batas bawah dan atas berdasarkan 1.5 kali rentang IQR, metode ini memberikan cara standar untuk menentukan apakah data terlalu jauh dari pusat distribusi. metode ini tidak memerlukan asumsi tertentu tentang distribusi data, sehingga dapat diterapkan secara luas pada berbagai jenis dataset. Hal ini membuatnya menjadi pilihan yang sederhana namun andal untuk membersihkan data dari pengaruh nilai-nilai yang menyimpang.

Setelah outlier berhasil dihapus, didapatkan hasil visualisasi Scatter Plot dan Heatmap setelah outlier dihapus sebagai berikut :







Dari visualisasi diatas bisa dilihat jika proses penghapusan outlier memberikan dampak yang besar terhadap data yang akan kita digunakan. Sebagai contoh pada Scatter Plot, kini nilai maksimal dari jumlah kamar tidur yang ada hanya kira - kira 5 kamar tidur, begitu pula dari luas tanah dan juga luas bangunan yang memiliki variasi data yang dipersempit untuk menghilangkan / meminimalisir outlier. Dari heatmap juga bisa dilihat bahwa kini luas bangunan yang memiliki korelasi yang dekat dengan harga, dan luas tanah juga tidak berbeda jauh dengan luas bangunan dalam korelasinya dengan harga. Ini berarti nilai dari luas bangunan dan juga luas tanah berdampak signifikan terhadap harga rumah.

Berikut adalah tahapan dan perintah untuk menangani data outlier :

1. Membuat fungsi untuk melihat data outlier pada price, luas tanah, luas bangunan, dan kamar tidur

```r
# Menggunakan IQR untuk menghapus outlier dari kolom 'price'
# Fungsi untuk mendeteksi outlier menggunakan IQR
detect_outliers_IQR <- function(df, column_name) {
  Q1 <- quantile(df[[column_name]], 0.25)
  Q3 <- quantile(df[[column_name]], 0.75)
  IQR_value <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value

  # Menyaring data outlier
  outliers <- df %>%
    filter(df[[column_name]] < lower_bound | df[[column_name]] > upper_bound)

  return(outliers)
}

cat("Jumlah outlier pada price:", nrow(detect_outliers_IQR(data_main, "price")), "\n")
cat("Jumlah outlier pada luas_tanah:", nrow(detect_outliers_IQR(data_main, "luas_tanah")), "\n")
cat("Jumlah outlier pada luas_bangunan:", nrow(detect_outliers_IQR(data_main, "luas_bangunan")), "\n")
cat("Jumlah outlier pada kamar_tidur:", nrow(detect_outliers_IQR(data_main, "kamar_tidur")), "\n")
```

```
Jumlah outlier pada price: 816
Jumlah outlier pada luas_tanah: 617
Jumlah outlier pada luas_bangunan: 438
Jumlah outlier pada kamar_tidur: 201
```

2. Menghapus data yang terdeteksi sebagai outlier

```r
# Fungsi untuk mendeteksi dan menghapus outlier menggunakan IQR
remove_outliers_IQR <- function(df, column_name) {
  # Menghitung Q1, Q3, dan IQR
  Q1 = quantile(df[[column_name]], 0.25)
  Q3 = quantile(df[[column_name]], 0.75)
  IQR_value = Q3 - Q1
  lower_bound = Q1 - 1.5 * IQR_value
  upper_bound = Q3 + 1.5 * IQR_value

  # Menghapus data yang berada di luar rentang batas bawah dan atas
  df_no_outliers = df %>%
    filter(df[[column_name]] >= lower_bound & df[[column_name]] <= upper_bound)

  return(df_no_outliers)
}

data_clean = remove_outliers_IQR(data_main, "price")
data_clean = remove_outliers_IQR(data_clean, "luas_tanah")
data_clean = remove_outliers_IQR(data_clean, "luas_bangunan")
data_clean = remove_outliers_IQR(data_clean, "kamar_tidur")
```

3. Karena data yang memiliki harga > 10m sangat sedikit, dan data yang < 500jt kebanyakan tidak valid, maka disini kita akan menghapus semua data yang memiliki harga < 500jt dan > 10m

```r
# Menentukan nilai batas harga yang terlalu kecil
batas_harga_min = 500

# Menentukan nilai batas harga yang terlalu Besar
batas_harga_max = 10000

# Menghapus data dengan harga diluar range
data_clean = data_clean %>%
  filter(price >= batas_harga_min) %>%
  filter(price < batas_harga_max)
```

Dari semua proses diatas, diperoleh data yang siap untuk dilanjutkan ke tahap modeling dengan summary data sebagai berikut :

1. Harga (dalam jutaan)

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  500    1400    2300    2760    3600    9900
```

2. Luas bangunan

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
16.0   100.0   150.0   163.6   210.0   450.0
```

3. Luas tanah

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
18.0    78.0   107.0   128.8   168.0   380.0
```

4. Kamar tidur

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.000   3.000   3.000   3.533   4.000   5.000
```

4. **Modelling & Evaluation**

Pada kasus ini kita menggunakan algoritma **Random Forest** yang mana merupakan algoritma **Supervised Learning** berbasis ensemble yang terdiri dari banyak decision trees. Dalam kasus ini, Random Forest digunakan untuk regresi guna memprediksi harga rumah berdasarkan fitur numerik (luas tanah, luas bangunan, jumlah kamar tidur) dan kategorikal (kota). Alasan kita menggunakan algoritma ini adalah karena algoritma ini sangat bagus dalam menangani data yang rumit, seperti hubungan antara fitur geografis dan harga rumah, yang seringkali tidak linear. Random Forest juga tahan terhadap kesalahan data (misalnya, ada harga yang sangat tinggi atau sangat rendah) dan mampu bekerja baik dengan data campuran, seperti angka (luas tanah) dan kategori (wilayah)[2].

Untuk langkah evaluasi, data dipecah menjadi dua bagian yaitu **data training (80%)** dan **data testing (20%)**. Data harga rumah kemudian ditransformasi dengan menggunakan logaritma untuk menangani potensi skala harga yang besar dan membantu model mempelajari pola dengan lebih baik. Selanjutnya, fitur numerik di-normalisasi menggunakan metode scaling untuk memastikan semua fitur berada pada skala yang sama, yang penting untuk memastikan model tidak bias terhadap fitur dengan skala yang lebih besar.

Transformasi data price

```{r}
data_model = data_clean
data_model$price = log(data_clean$price + 1)
```

Pisahkan data menjadi x(fitur) dan y(target)

```{r}
X = data_model[, c("kamar_tidur", "luas_tanah", "luas_bangunan", "kota")]
y = data_model$price
```

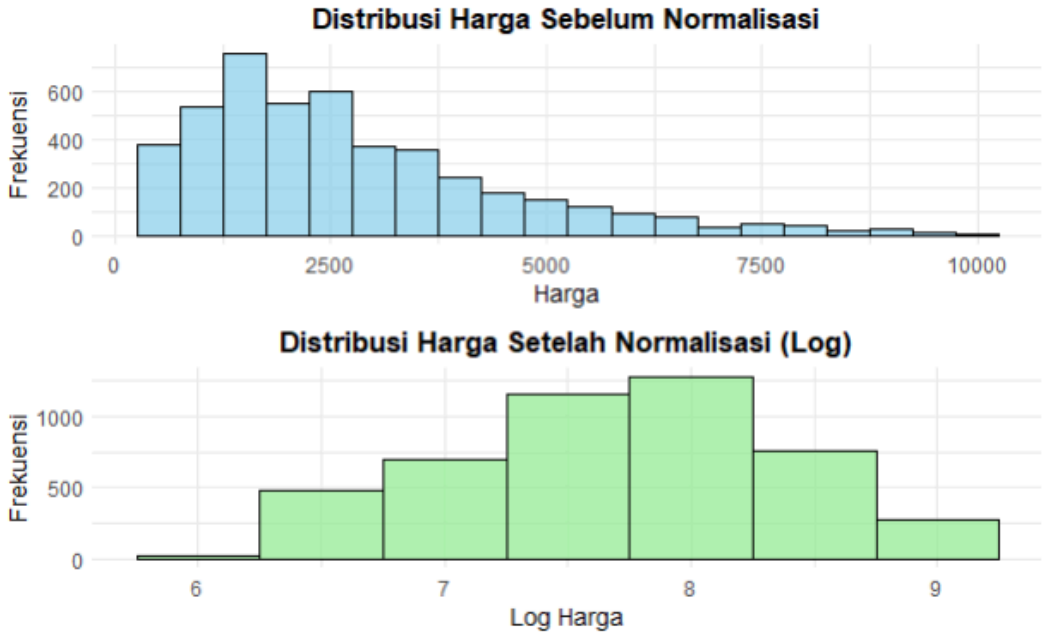
Membagi data menjadi dua (Training 80%, Testing 20%)

```{r}
set.seed(123)
trainIndex = createDataPartition(y, p = 0.8, list = FALSE)
train_data = data_model[trainIndex,]
test_data = data_model[-trainIndex,]
```

Menyiapkan data untuk Random Forest

```{r}
train_features = train_data[, c("kamar_tidur", "luas_tanah", "luas_bangunan", "kota")]
test_features = test_data[, c("kamar_tidur", "luas_tanah", "luas_bangunan", "kota")]
train_target = train_data$price
test_target = test_data$price
```

Berikut adalah visualisasi perbandingan data harga sebelum dan sesudah di-normalisasi :



Setelah itu, model dilatih menggunakan Random Forest dengan proses tuning parameter hyperparameter utama, yaitu "mtry" (jumlah fitur yang dipilih secara acak pada setiap split pohon). Model ini juga dievaluasi menggunakan validasi silang (cross-validation) sebanyak lima lipatan untuk memastikan performa yang stabil dan generalisasi yang baik.

Melakukan normalisasi fitur

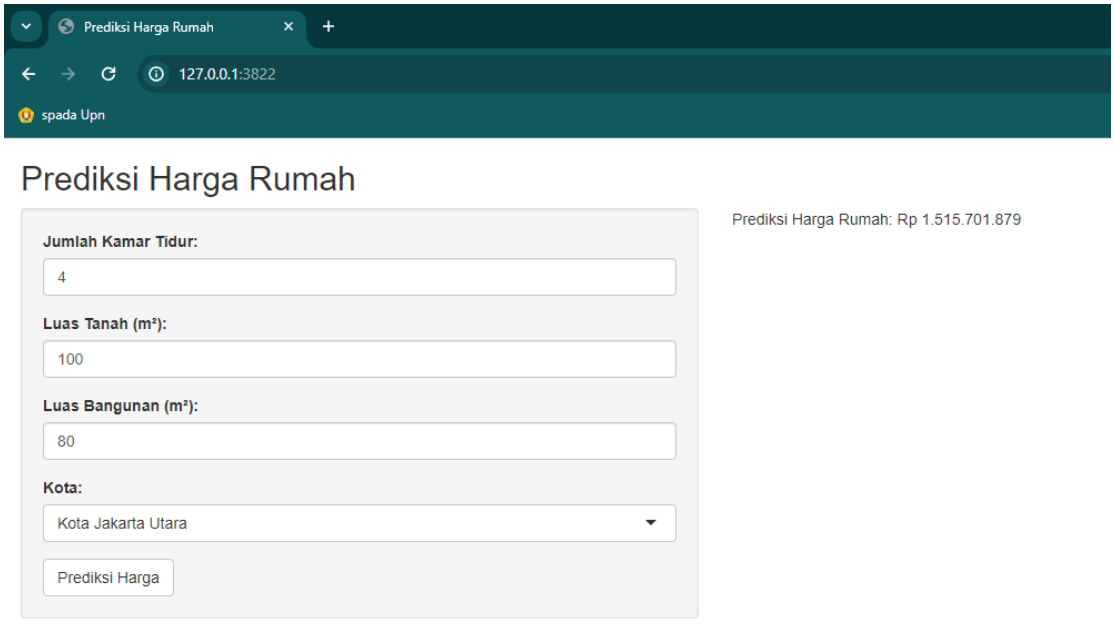
```{r}
pre_process = preProcess(train_features, method = "scale")
train_features_scaled = predict(pre_process, train_features)
test_features_scaled = predict(pre_process, test_features)
```

Tuning model

```{r}
tune_grid = expand.grid(.mtry = c(1, 2, 3, 4, 5))  # Tuning mtry
rf_tuned = train(
  price ~ kamar_tidur + luas_tanah + luas_bangunan + kota,
  data = train_data,
  method = "rf",
  tuneGrid = tune_grid,
  trControl = trainControl(method = "cv", number = 5))  # Cross-validation

saveRDS(rf_tuned, "rf_model_tuned.rds")
```

Contoh result model :



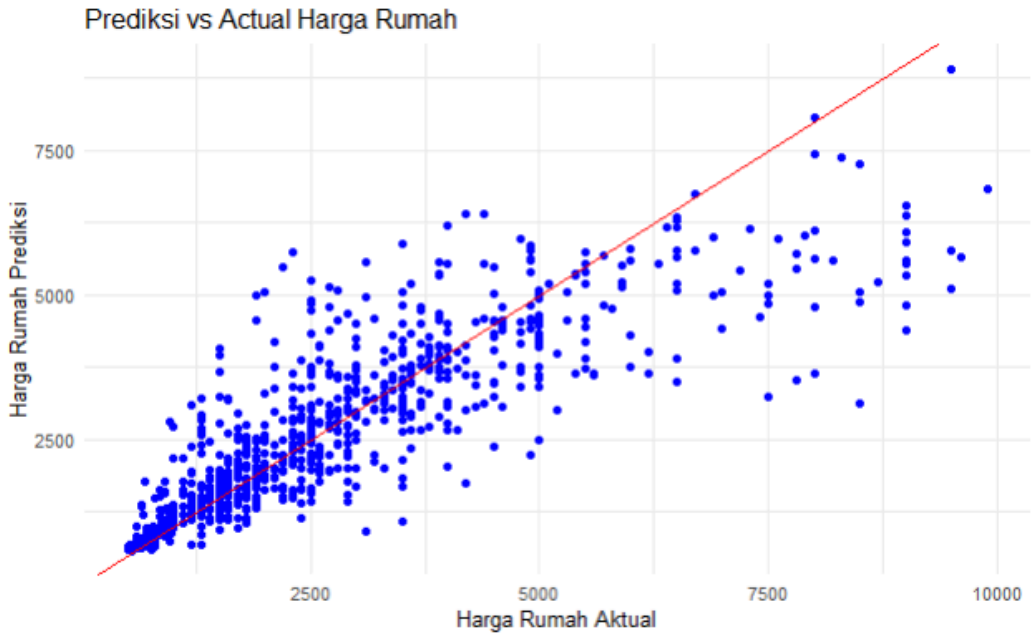
Setelah model ini dilatih menggunakan data yang kita miliki, kita mengevaluasinya untuk melihat seberapa baik prediksinya. Kita menggunakan beberapa cara untuk mengukur performanya, seperti:

- Seberapa besar rata-rata kesalahan prediksi (MAE)
- Seberapa besar model ini mampu menjelaskan variasi harga rumah (R^2)

Berikut hasil evaluasi yang diperoleh setelah proses modeling:

```
R-squared for Random Forest model:  0.8084741
RMSE for Random Forest model:  0.296078
MAE for Random Forest model:  0.2103679
RSE for Random Forest model:  0.2963979
MSE for Random Forest model:  0.08766219
```

Dari hasil evaluasi, model menunjukkan performa yang cukup baik dengan R-squared sebesar 0.8, yang artinya model mampu menjelaskan sekitar 80% variasi harga rumah. RMSE berada pada angka 0.29 (dalam skala log harga rumah), yang menunjukkan bahwa rata-rata kesalahan prediksi cukup kecil. MAE juga relatif rendah, yaitu 0.210, yang mengindikasikan bahwa kesalahan absolut rata-rata dalam prediksi log harga rumah juga kecil. Berikut adalah visualisasi perbandingan hasil prediksi dengan harga asli:



Pada visualisasi data diatas, diketahui bahwa semakin tinggi harga, akurasi semakin menurun. Setelah kita analisis, hal ini terjadi karena skala harga yang sangat beragam dimana dalam dataset, harga rumah memiliki rentang yang sangat besar, mulai dari harga rumah yang rendah hingga sangat tinggi. Model Random Forest, meskipun fleksibel, cenderung kesulitan menangkap pola yang jarang terjadi pada data, seperti outlier atau harga rumah yang jauh di atas rata-rata. Selain itu, distribusi data yang tidak seimbang dimana sebagian besar data memiliki harga rumah di level rendah atau menengah, model akan lebih fokus belajar dari pola pada rentang harga tersebut. Dengan kata lain, model cenderung *"bias"* ke harga yang lebih umum karena data yang merepresentasikan harga tinggi lebih sedikit. Akibatnya, untuk harga rumah tinggi, model mungkin tidak memiliki cukup informasi untuk membuat prediksi yang akurat.

5. Kesimpulan

Dalam proyek ini, kami telah mengidentifikasi permasalahan utama dalam pasar properti di Jakarta, yaitu ketidakpastian harga yang sering dihadapi oleh pembeli, penjual, dan investor. Untuk membantu mengatasi masalah ini, kami mengembangkan sebuah model prediksi harga properti menggunakan algoritma Random Forest Regressor. Model ini menunjukkan performa yang baik dengan nilai R^2 sebesar 0.8, yang berarti model mampu menjelaskan sekitar 80% variasi harga properti berdasarkan fitur seperti luas tanah, luas bangunan, jumlah kamar, dan lokasi. Hasil prediksi ini memberikan manfaat signifikan, terutama bagi pembeli yang dapat menggunakan estimasi harga untuk menilai kewajaran properti, serta bagi penjual dan investor untuk menetapkan harga yang lebih kompetitif dan informasional.

Meskipun model telah memiliki performa yang baik, terdapat beberapa hal yang bisa dioptimalkan untuk meningkatkan akurasi dan adaptabilitasnya. Akurasi cenderung menurun pada harga properti yang sangat tinggi, yang disebabkan oleh distribusi data yang tidak merata. Selain itu, model ini bergantung pada data historis, sehingga kurang adaptif terhadap perubahan mendadak di pasar properti. Oleh karena itu, diperlukan langkah-langkah tambahan, seperti memperluas cakupan data di rentang harga tertentu khususnya harga tinggi, serta melakukan segmentasi pasar untuk menghasilkan prediksi yang lebih spesifik pada setiap kategori. Selain itu mengintegrasikan hasil model ke dalam platform digital untuk mempermudah pengguna dalam membuat keputusan. Dengan strategi-strategi ini, model dapat mendukung terciptanya pasar properti yang lebih transparan dan efisien, memberikan manfaat yang signifikan bagi semua pihak yang terlibat.

*definisi bisnis yang dimaksud di bagian ini bukan wajib bidang bisnis, namun berarti berbagai bidang (sesuai yang telah dijelaskan/ dicoba implementasinya pada mini project)