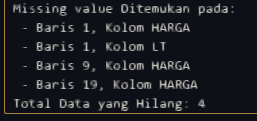
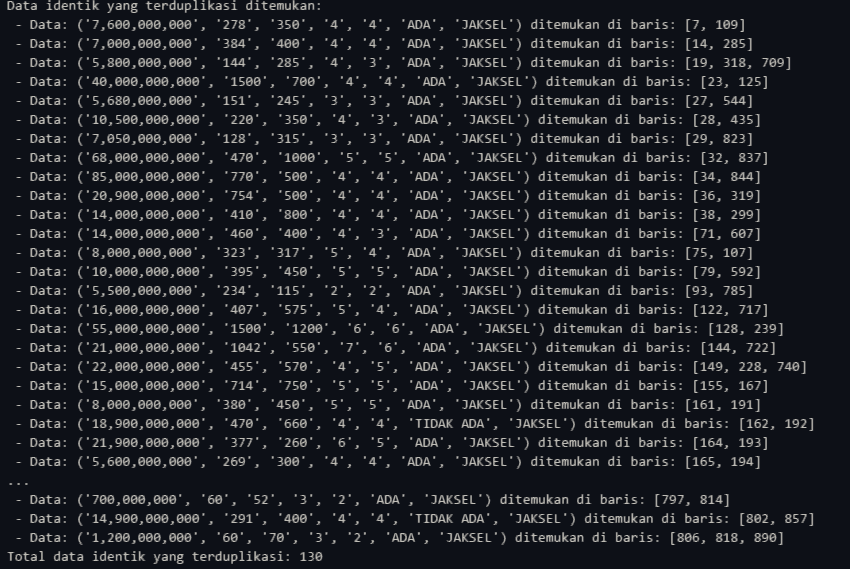
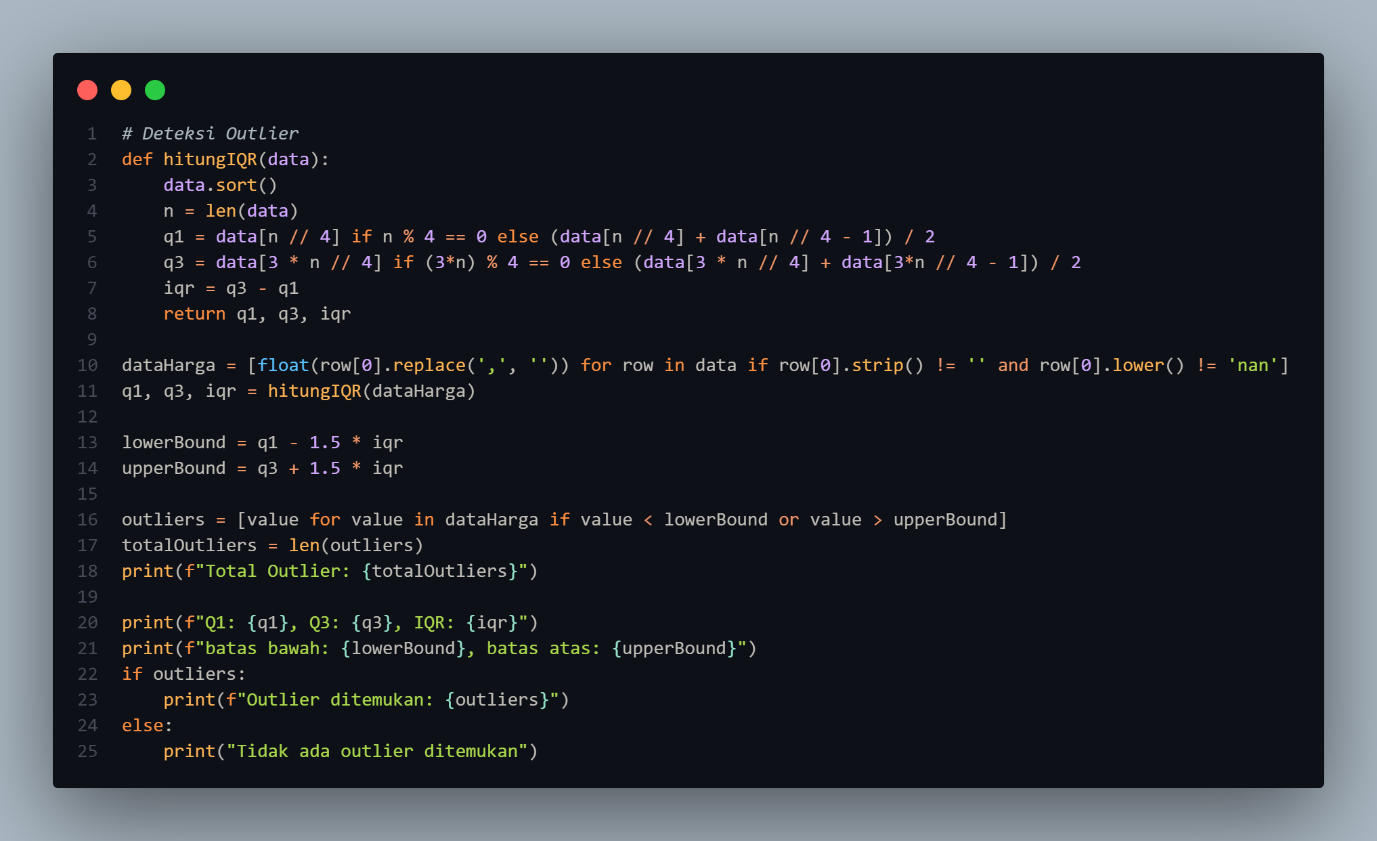
|  |  |
| --- | --- |
| Tugas Kelompok |  |
| Nama Kelompok | 1. Moh Naufal Faqih (10222044) 2. Firman Firdaus (10222033) 3. Ryan Azis Saputra (10222041) 4. Andika Fitra Ramadan (10222043) |
| Mata Kuliah | Pengolahan dan Analisis Sains Data |
| Dosen | Dr. Susmini Indriani Lestariningati, S.T, M.T. |
| Program Studi | Sistem Komputer |

1. **Deteksi Missing Value**
   1. **Program**
   2. **Output**



1. **Deteksi Duplikasi Data**
   1. **Program**
   2. **Output**
   3. **Catatan Tambahan**

* Deteksi pada duplikasi hanya dilakukan pada data yang benar-benar identik, dengan mempertimbangkan nilai pada kolom lainnya. Seperti luas tanah (LT), luas bangunan (LB), jumlah kamar tidur (JKT), jumlah kamar mandi (JKM), dan garasi (GRS).
* Jika deteksi duplikasi hanya dilakukan pada nilai kolom harga saja itu akan terdeteksi 735 data yang harganya mempunyai nilai yang sama.

1. **Deteksi Outlier Menggunakan Metode IQR**
   1. **Program**
   2. **Output**

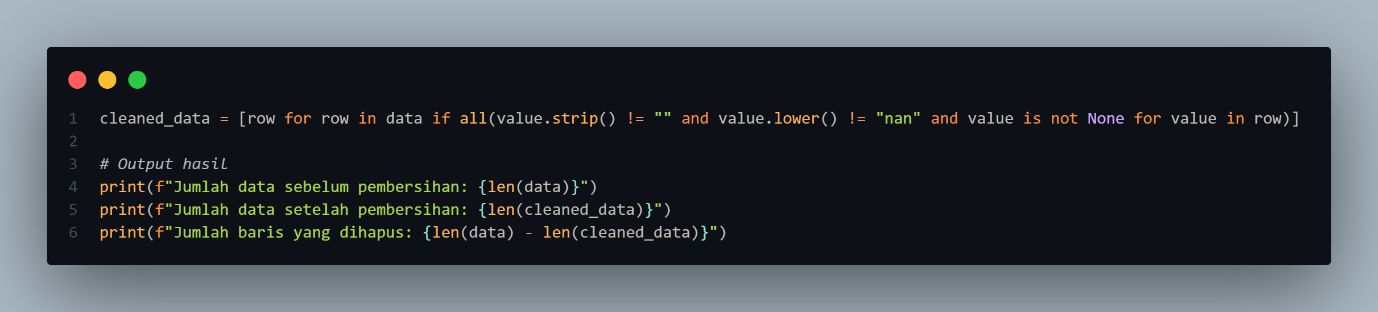
Total Outlier: 65

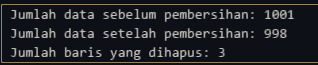
Q1: 6750000000.0, Q3: 20000000000.0, IQR: 13250000000.0

batas bawah: -13125000000.0, batas atas: 39875000000.0

Outlier ditemukan: [40000000000.0, 40000000000.0, 40000000000.0, 40000000000.0, 40000000000.0, 40000000000.0, 40000000000.0, 40000000000.0, 42000000000.0, 42000000000.0, 42000000000.0, 43000000000.0, 43000000000.0, 45000000000.0, 45000000000.0, 45000000000.0, 47000000000.0, 49000000000.0, 50000000000.0, 50000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 55000000000.0, 56000000000.0, 57000000000.0, 60000000000.0, 60000000000.0, 60000000000.0, 60000000000.0, 65000000000.0, 65000000000.0, 68000000000.0, 68000000000.0, 68000000000.0, 68000000000.0, 72000000000.0, 75000000000.0, 75000000000.0, 75000000000.0, 77500000000.0, 80000000000.0, 85000000000.0, 85000000000.0, 85000000000.0, 85000000000.0, 85500000000.0, 90000000000.0, 95000000000.0, 100000000000.0, 110000000000.0, 120000000000.0, 135000000000.0, 165000000000.0, 168000000000.0, 169000000000.0, 175230000000.0, 180000000000.0, 185000000000.0, 250000000000.0]

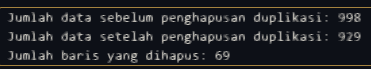
* 1. **Catatan Tambahan:**
* Total outlier sebanyak 65 data terdeteksi sebagai outlier.
* Ini menunjukkan bahwa ada banyak data yang berada di luar batas distribusi normal harga rumah.
* Q1 (Kuartil Pertama) 6.75 miliar (6.750.000.000.0) menunjukkan 25% dari data memiliki harga rumah di bawah nilai ini.
* Q3 (Kuartil Ketiga) 20 miliar (20.000.000.000.0) menunjukkan 75% dari data memiliki harga rumah di bawah nilai ini.
* IQR (Interquartile Range) 13.25 miliar (13.250.000.000.0) menunjukkan rentang variasi harga rumah antara Q1 dan Q3 di tengah distribusi.
* Harga rumah yang sangat tinggi (di atas 39.875 miliar) kemungkinan besar mencerminkan properti mewah atau lokasi premium dengan mempertimbangkan faktor faktor lainnya, seperti luas tanah, luas bangunan, jumlah kamar tidur, jumlah kamar mandi, dan garasi.

1. **Proses Menangani Missing Value**
   1. **Program**
   2. **Output**

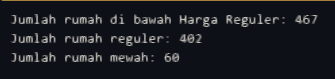


* 1. **Catatan Tambahan**
* Karena data yang missing/hilang hanya sedikit yaitu 3, melakukan penghapusan pada data adalah keputusan yang tepat dibanding melakukan mengisi (imputasi) yang bisa menambah bias atau error.
* Karena data hanya 0.3% yang dihapus dari seluruh data maka tidak akan memengaruhi rata-rata, median maupun sebaran data.

1. **Proses Menangani Duplikasi Data**
   1. **Program**
   2. **Output**



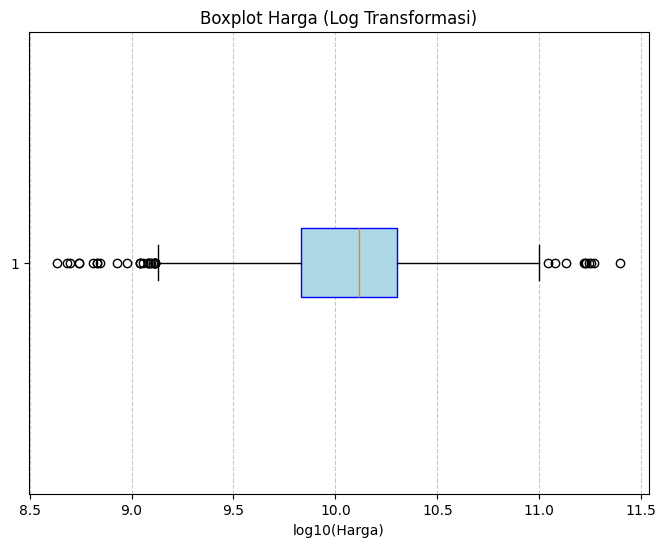
1. **Proses Menangani Outlier**
   1. **Program**
   2. **Output**

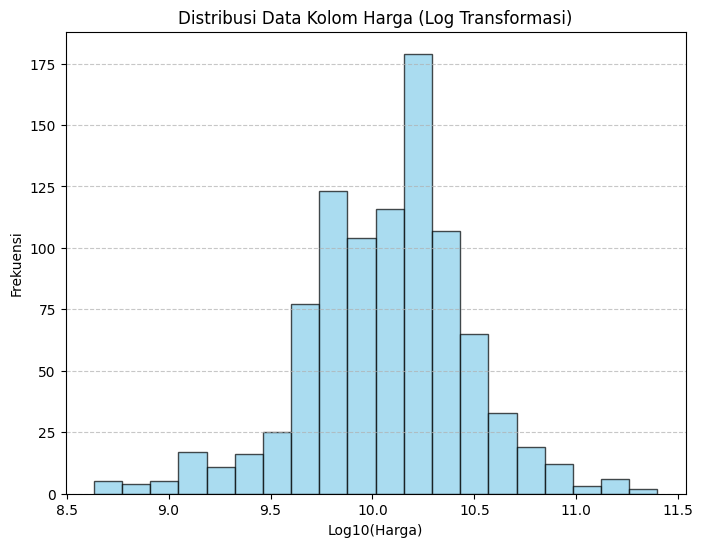


* 1. **Catatan Tambahan**
* Outlier pada data ditandai dengan membaginya ke beberapa segmen berdasarkan harganya (misalnya: di bawah harga reguler dan mewah).
* Ada beberapa rumah dengan harga sangat tinggi yang kemungkinan memiliki fasilitas eksklusif dan lengkap.
* Outlier tidak dianggap sebagai “kesalahan” atau “anomali”, tapi dijadikan kategori khusus yang memberi makna baru dalam analisis.
* Dengan mengelompokkan harga ekstrem ke segmen tersendiri, dapat menghindari distorsi pada analisis rata-rata, atau median, karena outlier tak lagi mempengaruhi data utama

1. **Visualisasi Data (Boxplot dan Histogram)**
   1. **Program**



* 1. **Output**



* 1. **Catatan Tambahan**
* Nilai pada harga rumah awalnya condong ke kiri, artinya Sebagian besar harga ada di level rendah-menengah tapi ada segelintir yang sangat mahal.
* Maka dari itu dilakukanlah transformasi logaritmik (misal: log10(harga)) menormalkan distribusi, membuatnya lebih mendekati distribusi normal sehingga lebih mudah dianalisis secara statistik.
* Tanpa log transform, data mahal bisa menutupi variasi data di harga rendah.
* Dengan log, semua nilai bisa terlihat lebih proporsional, sehingga struktur sebaran data menjadi lebih jelas (seperti kelompok harga tertentu atau outlier)
* Log transformasi menekan nilai-nilai ekstrem (contohnya rumah 200 miliar) sehingga outlier tidak terlalu mendistorsi skala grafik.