KELAS•WORK
by KELAS.COM

Housing Price

Muhammad Abil Khoiri

19 Januari, 2026

Latar Belakang Masalah

Analisis harga rumah penting dilakukan demi memahami pola pasar properti serta membantu pengambilan keputusan berbasis data yang akurat.

Melalui studi kasus Housing Price, proses Exploratory Data Analysis digunakan untuk mengeksplorasi karakteristik data, mendeteksi pola, distribusi, serta hubungan antar variabel numerikal dan kategorikal, sehingga visualisasi satu dimensi dan dua dimensi dapat memberikan insight yang mendukung analisis harga rumah secara komprehensif.

Tujuan Project

- Memahami karakteristik dataset Housing Price
- Menganalisis variabel numerikal dan kategorikal
- Mengidentifikasi pola, hubungan, dan anomali data

Metodologi

Overview:

- Load data & hapus nilai NaN
- Pisahkan data numerikal dan kategorikal
- Visualisasi 1D dan 2D
- Log-transform untuk data numerikal

Missing Value Handling

Persentase missing value tinggi pada fitur tertentu ditangani melalui imputasi, penghapusan selektif, atau pengelompokan kategori khusus agar kualitas data tetap terjaga dan mendukung analisis statistik serta pemodelan selanjutnya lebih akurat

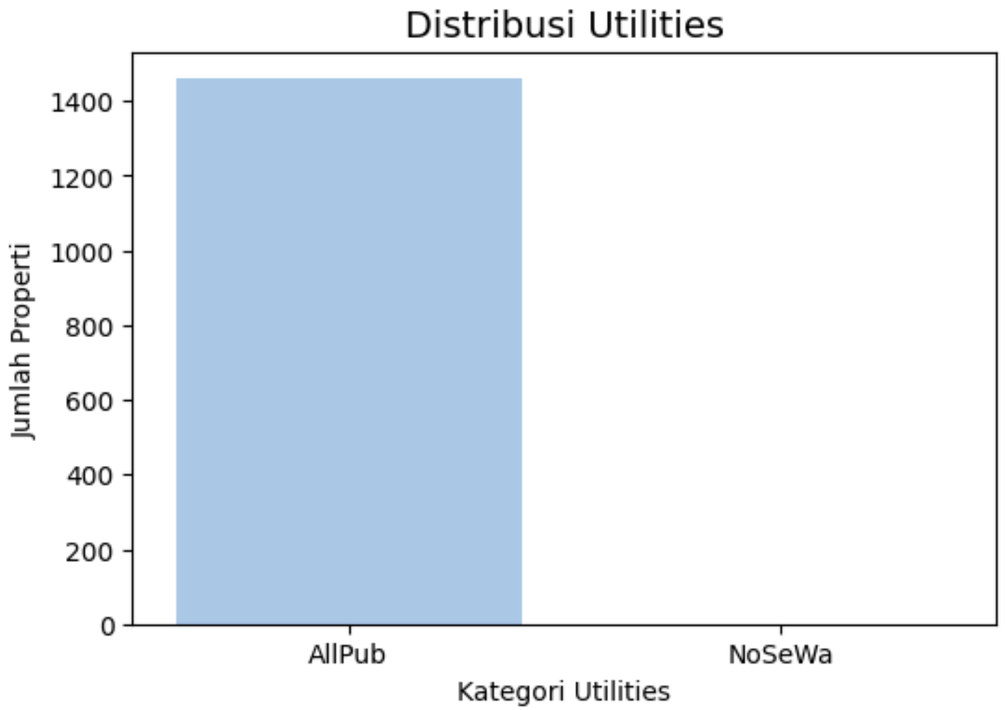
Penanganan Kolom Missing Tinggi
Kolom dengan missing value di atas dua puluh persen dipertimbangkan dihapus, namun dianalisis dahulu karena NaN bisa merepresentasikan ketiadaan fitur.

Imputasi Garage dan Basement
Kolom Garage dan Basement dengan missing rendah diimputasi None agar model mengenali rumah tanpa fasilitas tersebut secara eksplisit lebih akurat.

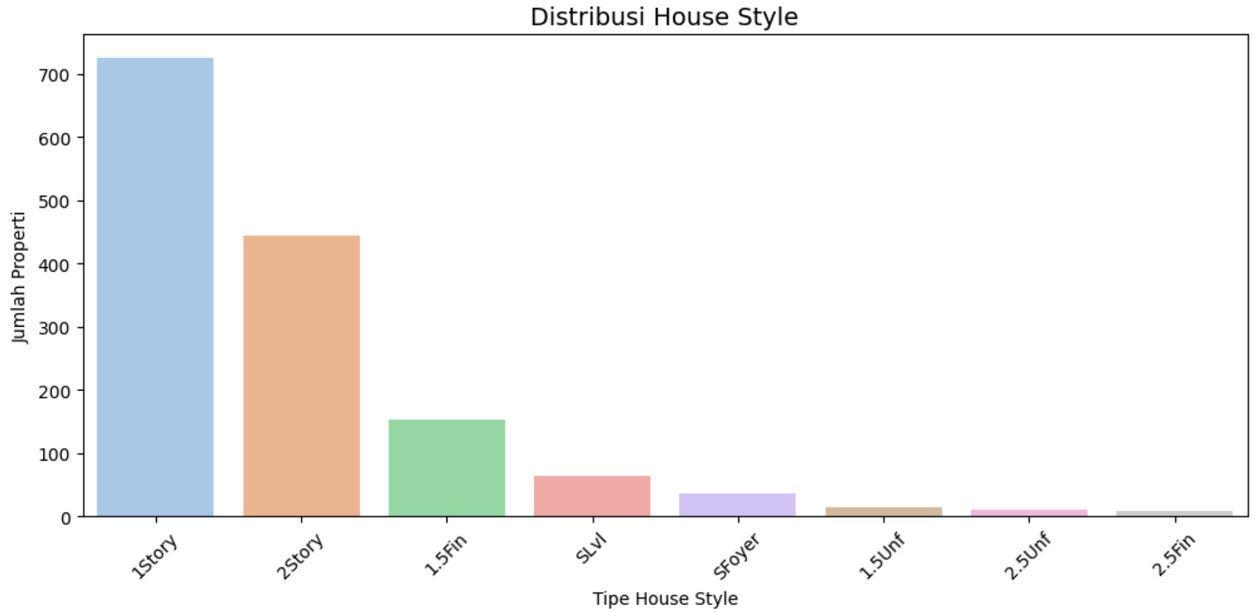
Imputasi Kolom Lainnya
Kolom lainnya ditangani menggunakan imputasi mean, median, atau mode berdasarkan tipe data untuk menjaga konsistensi dan kualitas dataset analisis lanjutan.

Distribusi Utilities

- Dominasi Mutlak: Hampir 100% properti memiliki fasilitas publik lengkap (AllPub), menunjukkan standar infrastruktur yang sangat merata.
- Keterbatasan Data: Kategori NoSeWa hampir tidak ada, menjadikannya variabel konstan yang tidak memberikan pengaruh variasi pada harga.



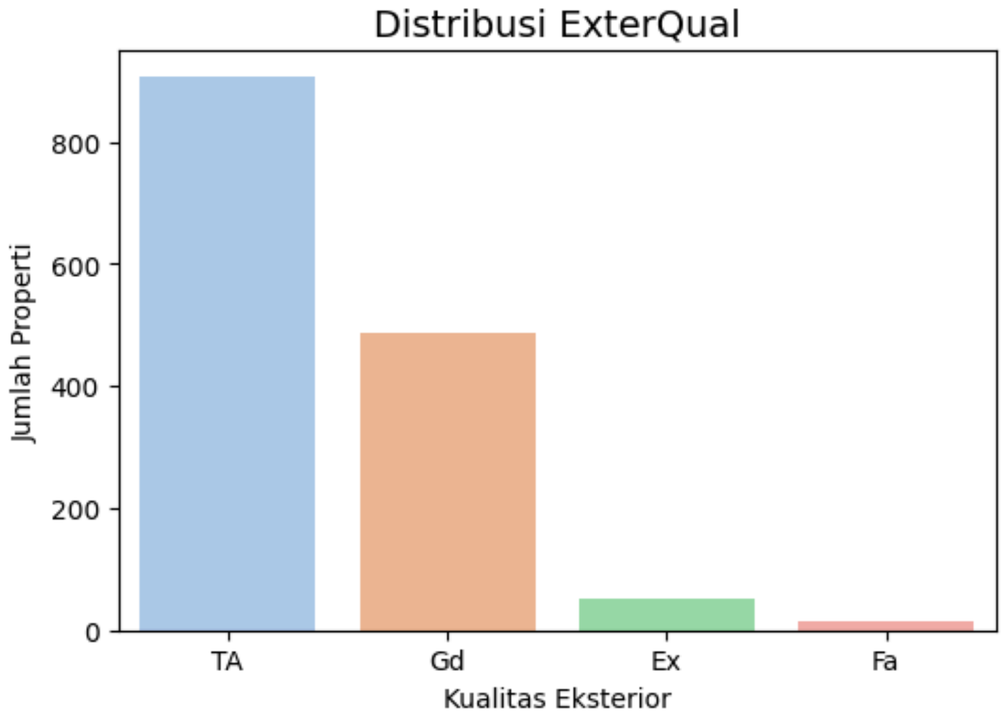
Distribusi House Style



- Tren Pasar: Mayoritas hunian adalah tipe 1Story (satu lantai) dan 2Story (dua lantai).
- Segmentasi: Desain konvensional mendominasi, sementara tipe seperti 2.5Fin atau SFoyer merupakan segmen yang sangat kecil/langka.

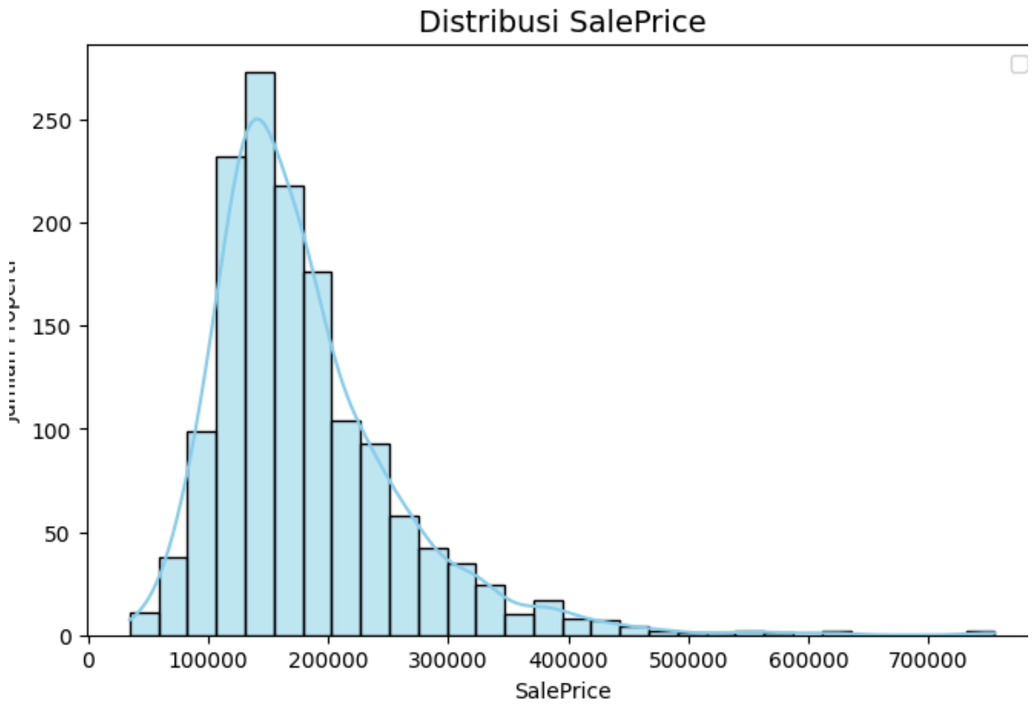
Distribusi ExterQual

- Dominasi Pasar: Lebih dari 90% properti menggunakan material eksterior dengan kualitas rata-rata (TA) hingga baik (Gd).
- Kelangkaan: Properti dengan kualitas mewah (Excellent) atau di bawah standar (Fair) merupakan anomali atau segmen pasar yang sangat kecil.
- Potensi Harga: Karena mayoritas berada di kategori Average, properti dengan status Good atau Excellent kemungkinan besar memiliki korelasi positif yang kuat terhadap kenaikan harga jual.



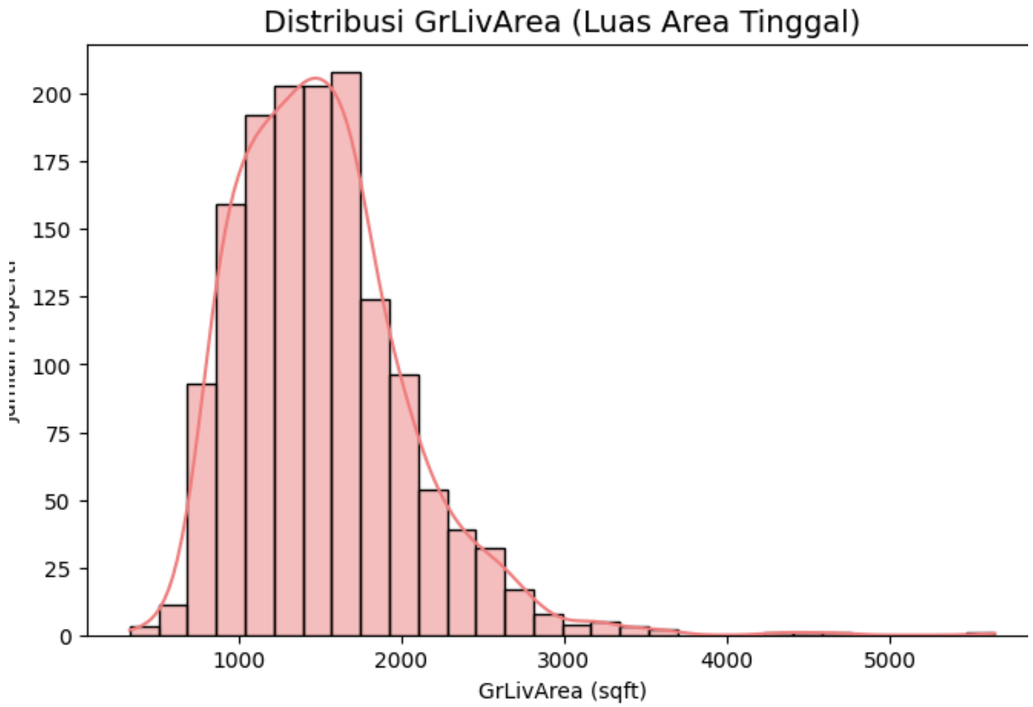
Distribusi SalePrice

- Distribusi Right-Skewed: Rata-rata harga (Mean) lebih tinggi dari nilai tengah (Median), menunjukkan adanya beberapa properti mewah yang sangat mahal.
- Variasi Harga Tinggi: Standar deviasi yang besar mencerminkan rentang harga properti yang sangat beragam di pasar.
- Konsentrasi Pasar: Mayoritas transaksi terfokus pada rumah dengan harga menengah (100k - 200k), sementara rumah kategori high-end sangat sedikit.



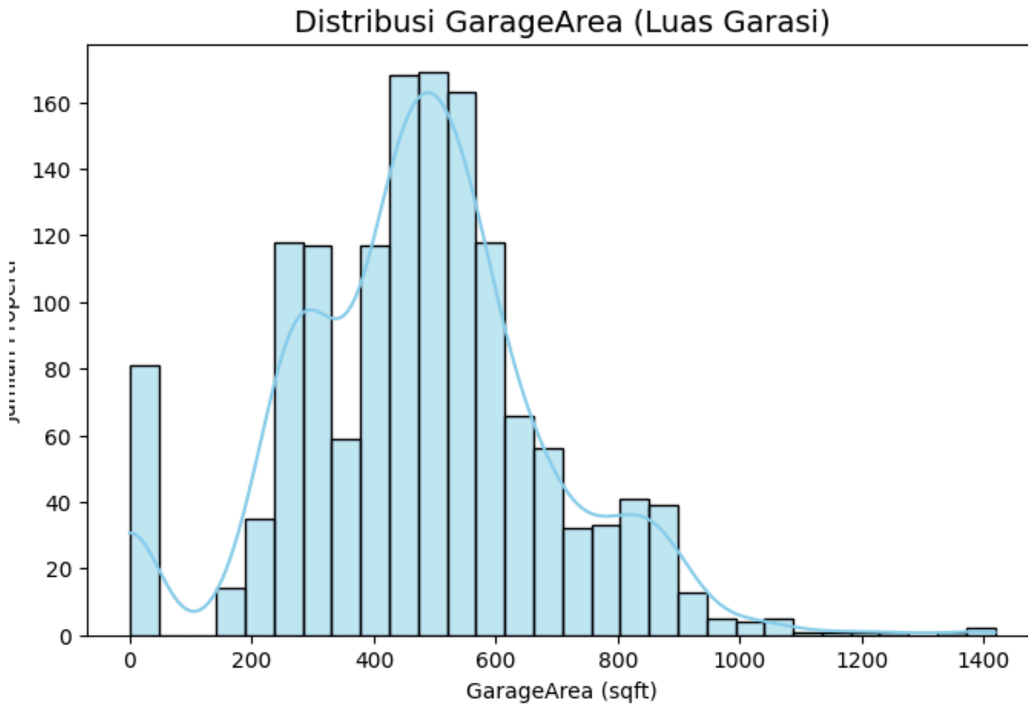
Distribusi GrLivArea

- Distribusi Right-Skewed: Rata-rata luas lebih besar dari median, menandakan adanya segelintir properti dengan luas area yang sangat besar.
- Variasi Luas Tinggi: Standar deviasi yang mencapai 525 sqft menunjukkan rentang ukuran rumah yang sangat beragam di pasar.
- Standar Pasar: Mayoritas properti terkonsentrasi pada luas 1.000 – 2.000 sqft, sementara rumah di atas 4.000 sqft merupakan kasus langka (outliers).

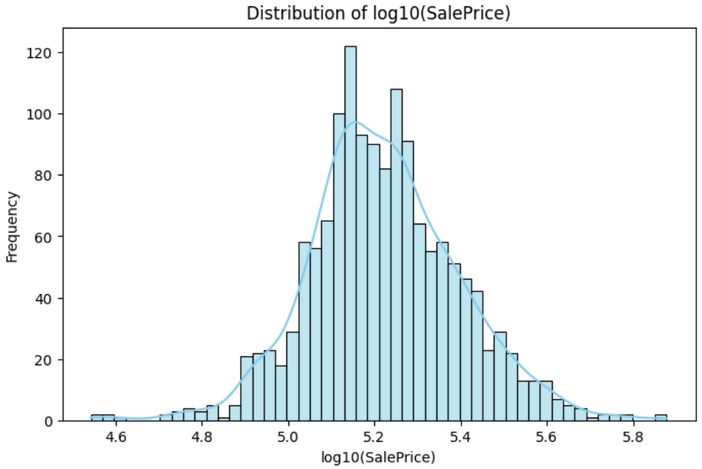


Distribusi GarageArea

- Distribusi Simetris: Nilai rata-rata dan median hampir identik, menunjukkan persebaran data yang seimbang dan konsisten.
- Variasi Ukuran: Standar deviasi yang tinggi mencerminkan perbedaan kapasitas, dari garasi kecil hingga kapasitas 3 mobil.
- Properti Tanpa Garasi: Adanya nilai nol menunjukkan sekelompok properti yang tidak memiliki fasilitas parkir tertutup.
- Standar Pasar: Mayoritas garasi berada di rentang 400–600 sqft, yang merupakan standar umum untuk dua mobil.

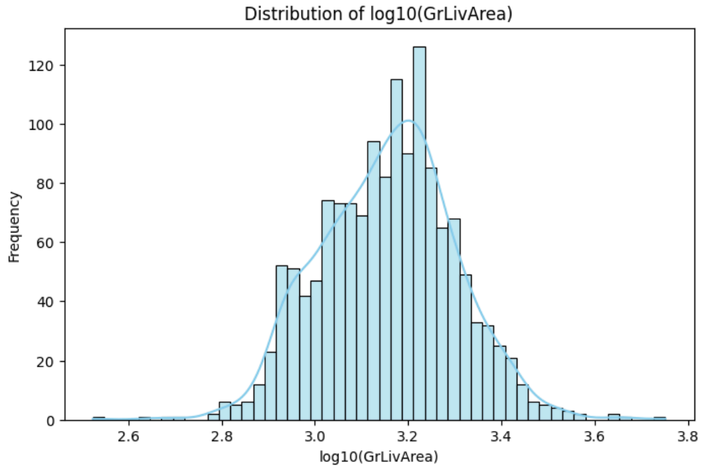


Normalisasi



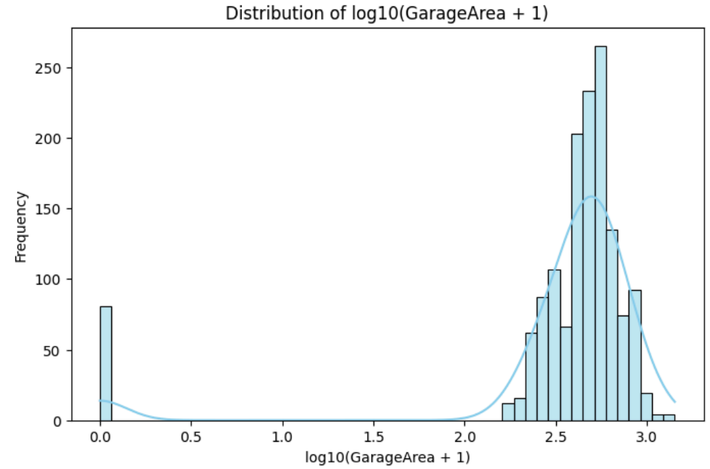
Optimalisasi SalePrice

- Normalitas: Transformasi berhasil menciptakan distribusi simetris dengan Mean (5.222) dan Median (5.212) yang hampir identik.
- Efek: Mengurangi skewness ekstrem, memastikan model prediksi tidak terdistorsi oleh harga rumah yang sangat mahal.



Normalisasi GrLivArea

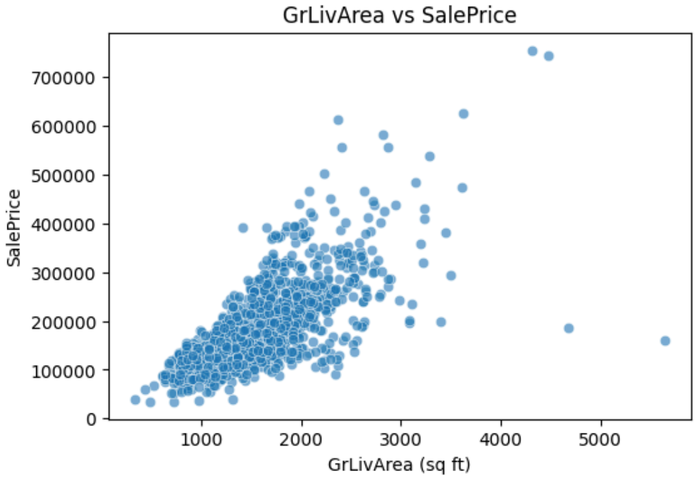
- Stabilitas Data: Menghasilkan distribusi normal (Mean ≈ Median) yang menekan dampak negatif dari pencilan luas bangunan ekstrim.
- Skalabilitas: Luas area dalam skala logaritmik mempermudah model dalam menangkap korelasi linier dengan harga.



Karakteristik GarageArea

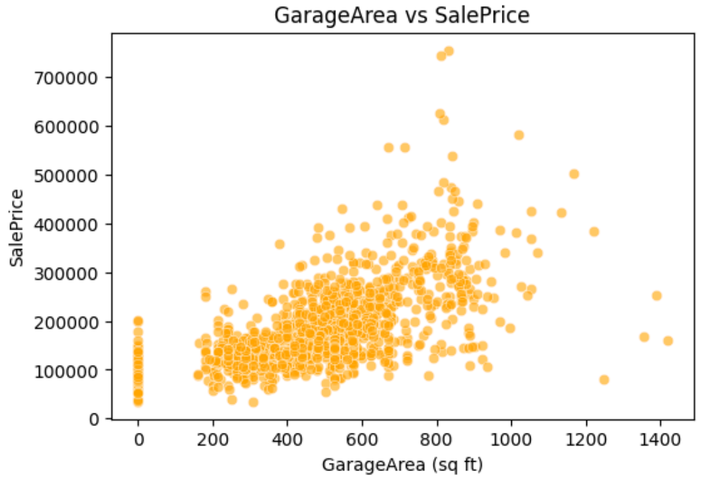
- Residu Skewness: Meskipun telah di-log, distribusi tetap miring ke kanan karena adanya nilai 0 (rumah tanpa garasi) yang cukup signifikan.
- Gap Fasilitas: Terdapat pemisahan jelas antara properti tanpa garasi (log=0) dan properti dengan garasi (log > 2.2).

GrLivArea Vs GarageArea



GrLivArea vs SalePrice

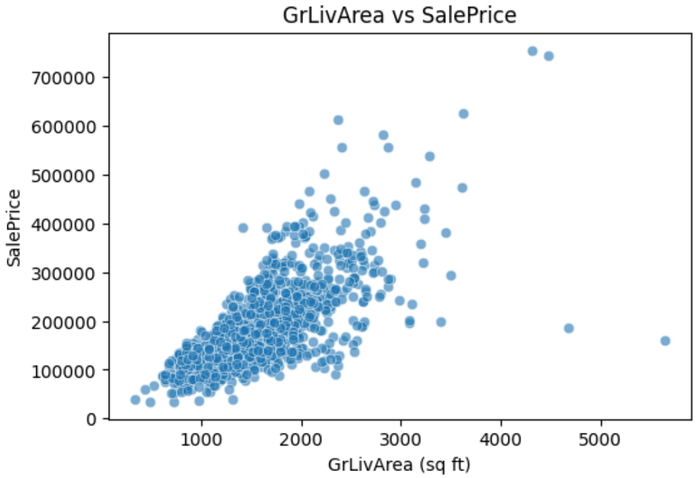
- Korelasi Positif Kuat: Grafik menunjukkan pola linear yang jelas, di mana peningkatan luas ruang tamu berbanding lurus dengan kenaikan harga jual.
- Faktor Penentu Utama: Sebagai bagian integral dari fungsi hunian, luas area tinggal menjadi prediktor harga yang paling konsisten dan padat.

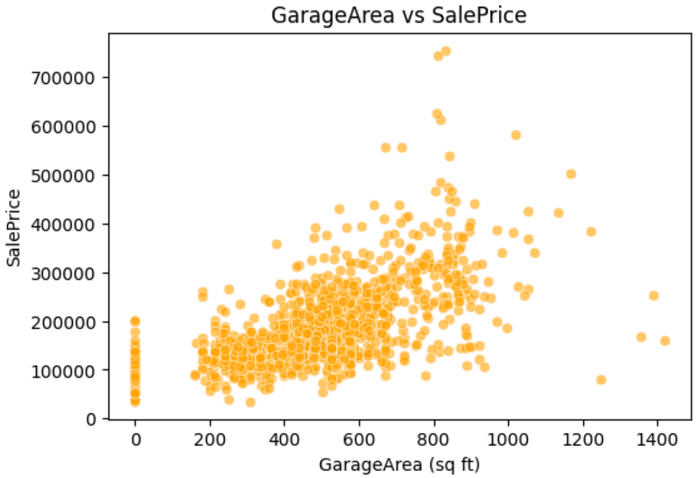


GarageArea vs SalePrice

- Fasilitas Tambahan: Terdapat korelasi positif, namun sebarannya lebih luas dibandingkan luas bangunan, menunjukkan pengaruh yang tidak terlalu dominan.
- Variansi Tinggi: Pada garasi luas (>1000 sqft), harga sangat bervariasi, mengindikasikan adanya faktor lain yang lebih menentukan nilai properti tersebut.

Perbandingan Pengaruh





- Signifikansi: Luas ruang tamu (GrLivArea) memiliki dampak yang jauh lebih signifikan terhadap harga jual dibandingkan luas garasi.
- Sinergi Mewah: Properti harga tertinggi biasanya merupakan kombinasi dari kedua variabel ini, di mana rumah mewah cenderung memiliki area tinggal sekaligus garasi yang besar.

Kesimpulan

- Pola & Anomali Data: EDA berhasil mengidentifikasi dominasi infrastruktur publik (AllPub) dan distribusi harga yang miring ke kanan (right-skewed).
- Variabel Penentu Utama: Luas area tinggal (GrLivArea) terbukti memiliki korelasi linear paling kuat terhadap kenaikan harga dibandingkan luas garasi.
- Kualitas Bangunan: Standar material eksterior rata-rata (TA) mendominasi pasar, sementara kualitas premium (Ex) menjadi pendorong harga yang signifikan.
- Kesiapan Modeling: Data telah melalui tahap Log Transformation untuk menormalkan distribusi, sehingga siap digunakan dalam model regresi prediktif.

02: Housing Price

https://github.com/MonyetttRindam/Project-Bootcamp-Kelas.com/tree/aaaee34b5e7541875b1b5cbcef4e4f71f5698b91/Modul%202

Thank You

19 Januari, 2026