

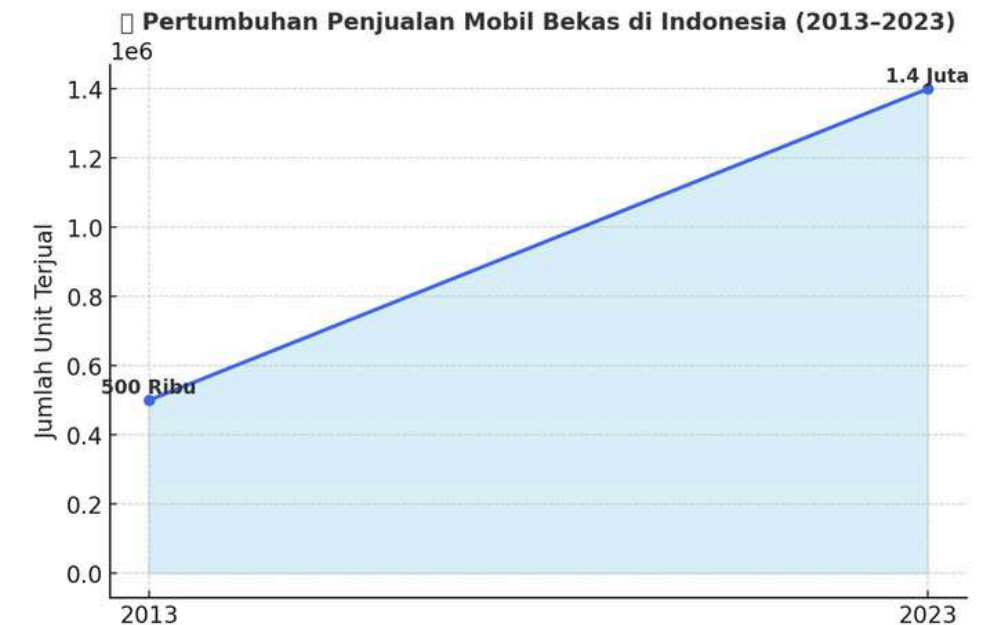
Final Project

Data Science

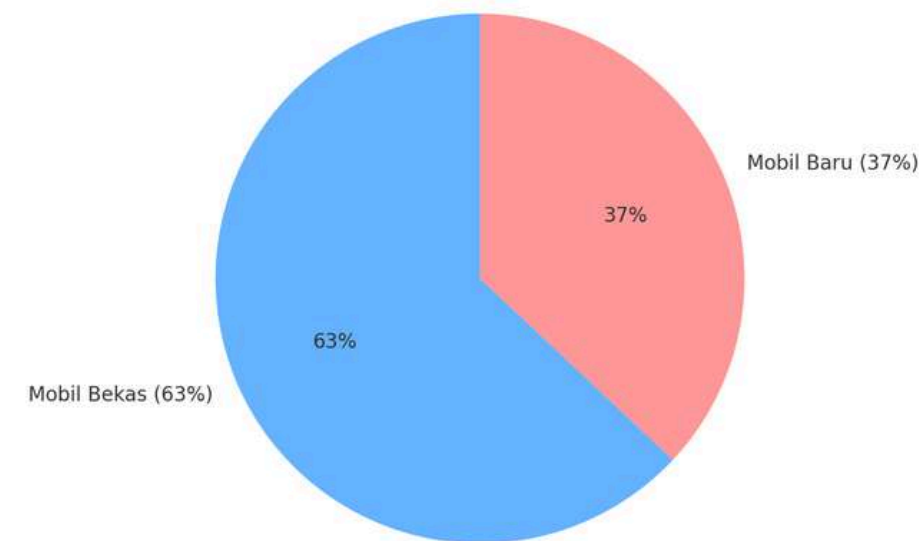
Alief Dhaffa (DS33B)

Problem Statement

- Penjualan mobil bekas di Indonesia **naik 180%** dalam 10 tahun terakhir, dari **500.000 unit (2013)** menjadi **1,4 juta unit (2023)**.
- **Preferensi bergeser.** Sekitar **63%** masyarakat di Jawa memilih mobil bekas dibanding mobil baru sepanjang 2023.
- Penentuan harga masih subjektif.



□ Preferensi Masyarakat Jawa dalam Membeli Mobil (2023)



Tujuan Bisnis:

1. Membangun model **prediksi harga mobil bekas** berdasarkan karakteristik mobil (tahun, km, fuel, transmission, ownership, dll).
2. Dengan model ini:
 - Penjual bisa menentukan harga jual yang kompetitif dan realistis.
 - Pembeli bisa mengetahui estimasi harga wajar sebelum membeli mobil bekas.
 - Platform jual-beli mobil (contoh: CarDekho, OLX, dll) dapat meningkatkan kepercayaan pengguna melalui transparansi harga.

Nilai Tambah dari Penyelesaian:

- **Meningkatkan kepercayaan pengguna** karena harga lebih transparan.
- **Mempercepat keputusan jual/beli** → memperbesar transaksi.
- **Mengurangi potensi sengketa harga** antara pembeli dan penjual.

The Data?



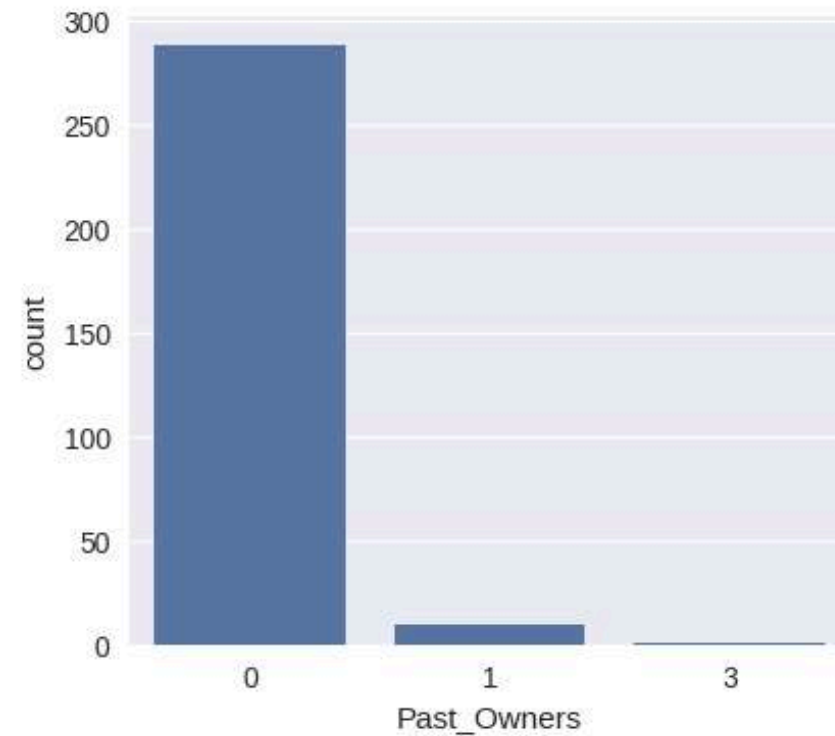
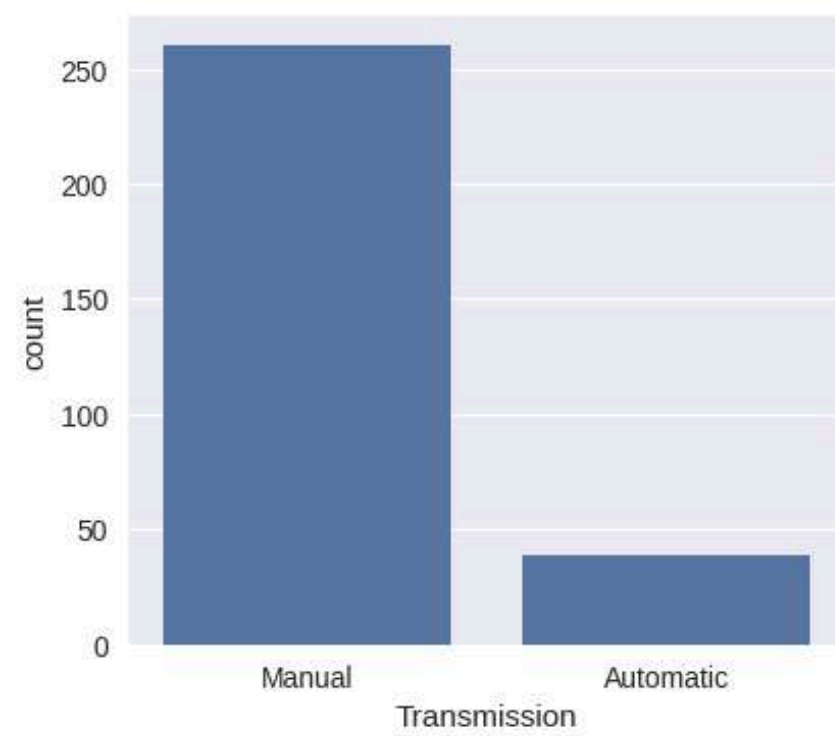
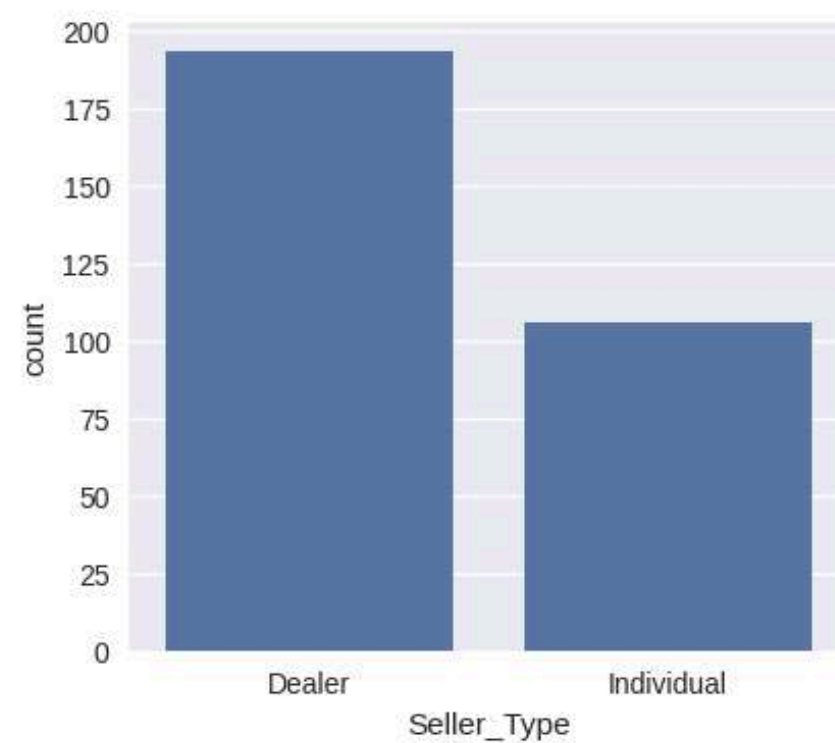
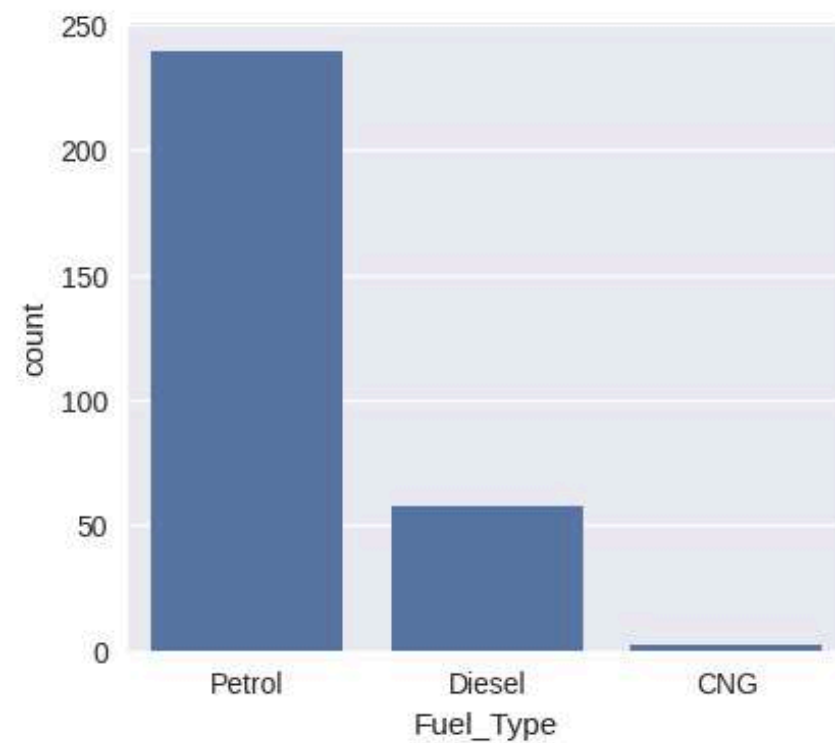
Vehicle dataset

kaggle

Dataset yang digunakan dalam proyek ini berasal dari CarDekho melalui Kaggle, berisi informasi 4.340 mobil bekas.

Kategori dan Distribusi:

- **name** → Nama mobil.
- **year** → Tahun produksi mobil.
- **selling_price** → Harga jual mobil (*target utama untuk prediksi*).
- **km_driven** → Jumlah kilometer yang sudah ditempuh.
- **fuel** → Jenis bahan bakar
- **seller_type** → Jenis penjual (Dealer, Individual, Trustmark Dealer).
- **transmission** → Jenis transmisi (Manual, Automatic).
- **owner** → Status kepemilikan (First Owner, Second Owner, etc).

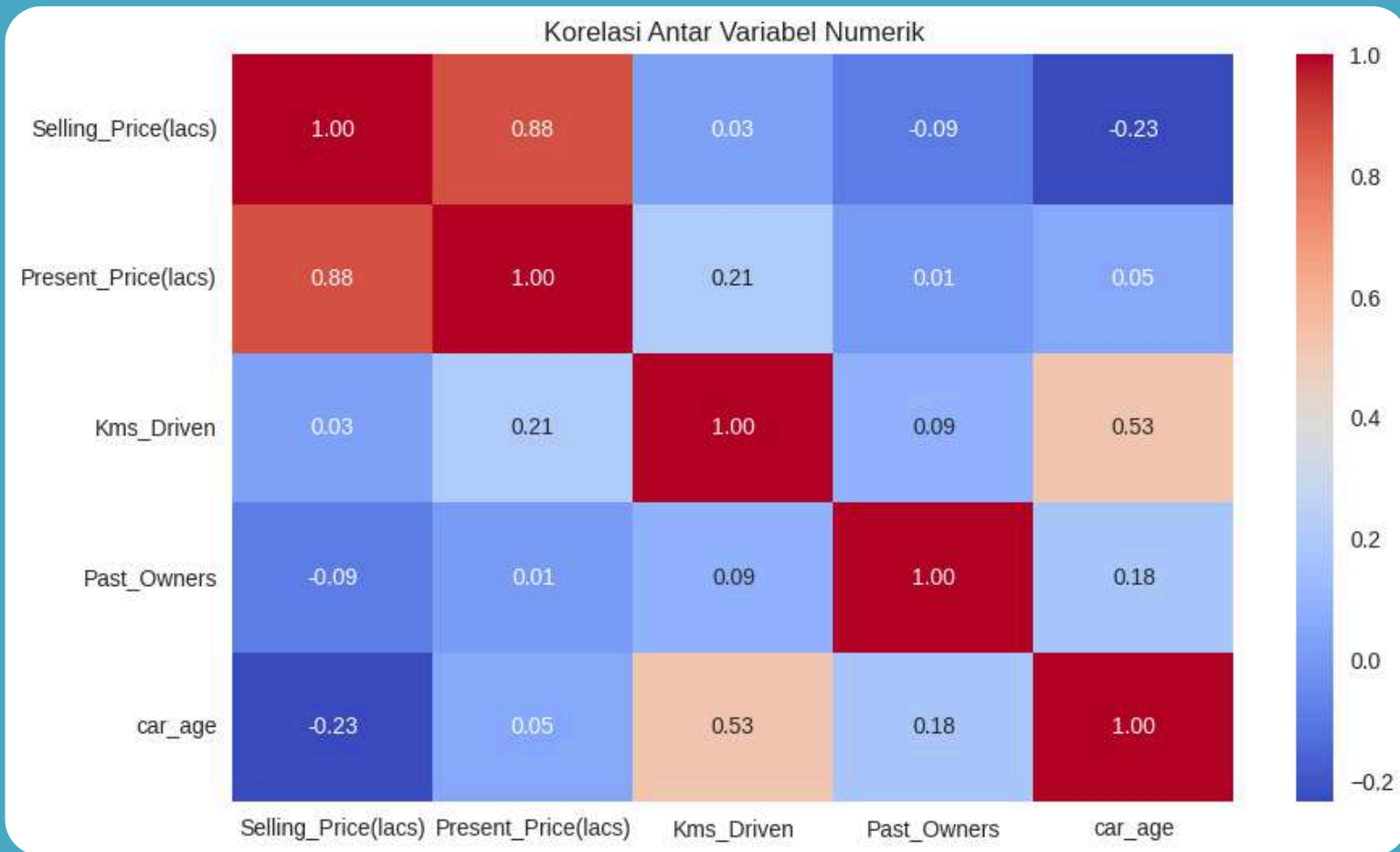


Petrol adalah jenis bahan bakar yang paling dominan, diikuti oleh Diesel dengan selisih signifikan, sementara CNG hampir tidak digunakan.

Dealer lebih banyak dibandingkan penjual individu, menunjukkan bahwa sebagian besar distribusi dilakukan oleh dealer.

Transmisi manual jauh lebih dominan dibandingkan transmisi otomatis.

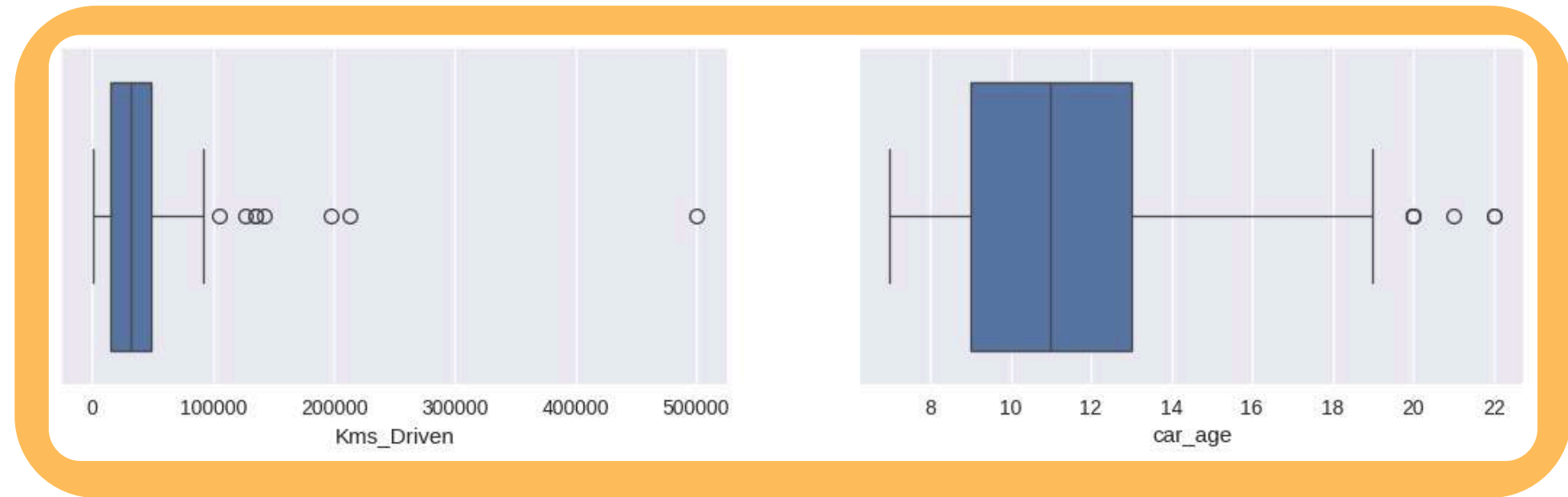
Sebagian besar kendaraan adalah kendaraan yang dulunya dibeli dari tangan pertama (0 pemilik sebelumnya), dengan sangat sedikit kendaraan yang memiliki pemilik sebelumnya.



▶ **Harga asli mobil sangat memengaruhi harga jual.**

▶ **Mobil tua cenderung lebih murah.**

▶ **Jumlah kilometer dan pemilik sebelumnya, memiliki pengaruh minimal, terhadap harga jual**



Banyak mobil dengan kilometer tinggi

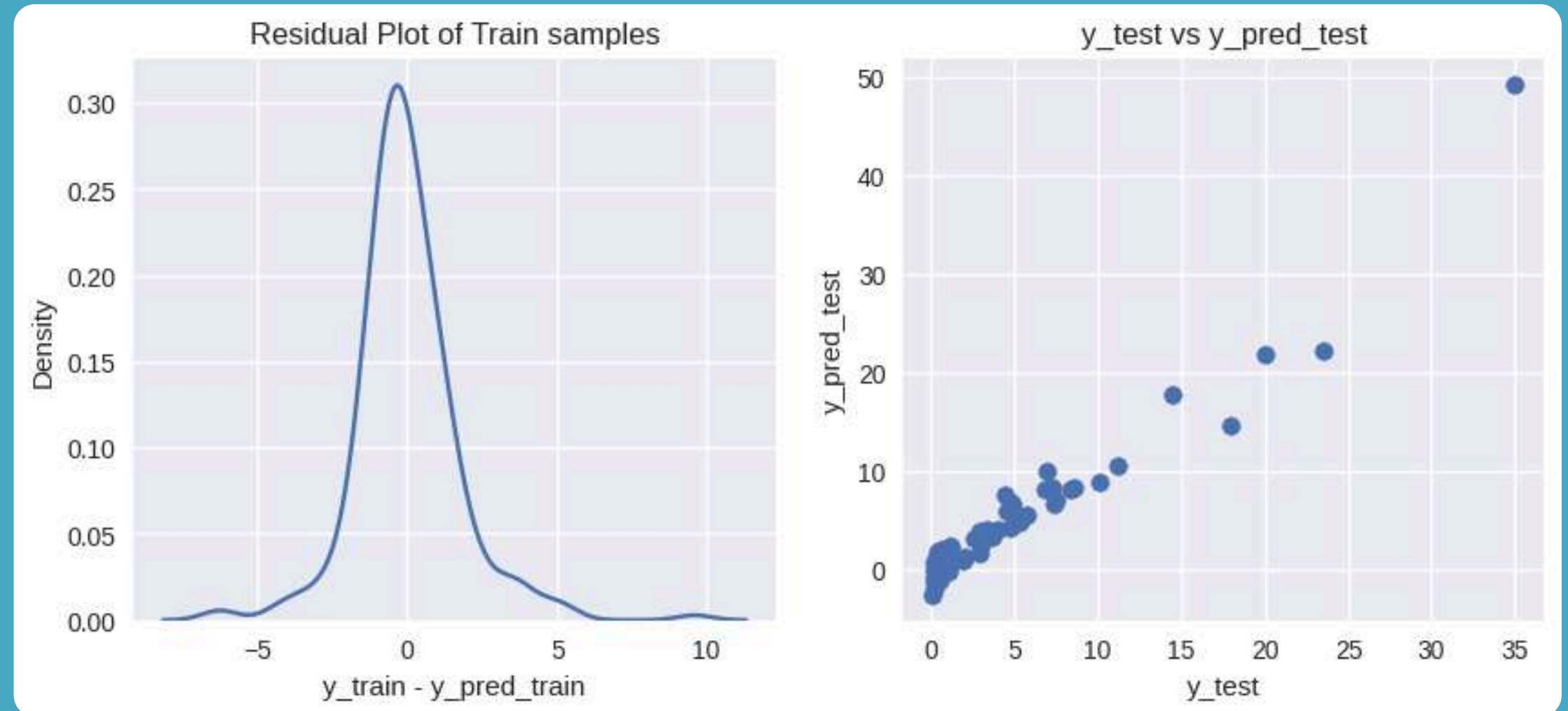
→ konsumen lebih suka mobil yang dianggap masih awet dan terjangkau.

Secara keseluruhan mobil berda dalam rentang usia 10-13 tahun

→ Pasar mobil bekas didominasi mobil yang masih dalam kurun waktu 10 tahun & jarak tempuh rendah

→ lebih menarik bagi pembeli.

STANDARD LINEAR REGRESSION



Nilai R^2 mendekati 1

→ model bisa menjelaskan 86–87% variasi harga mobil.

Train dan Test R^2 relatif seimbang
→ tidak ada overfitting (kalau train tinggi, test rendah baru bahaya).

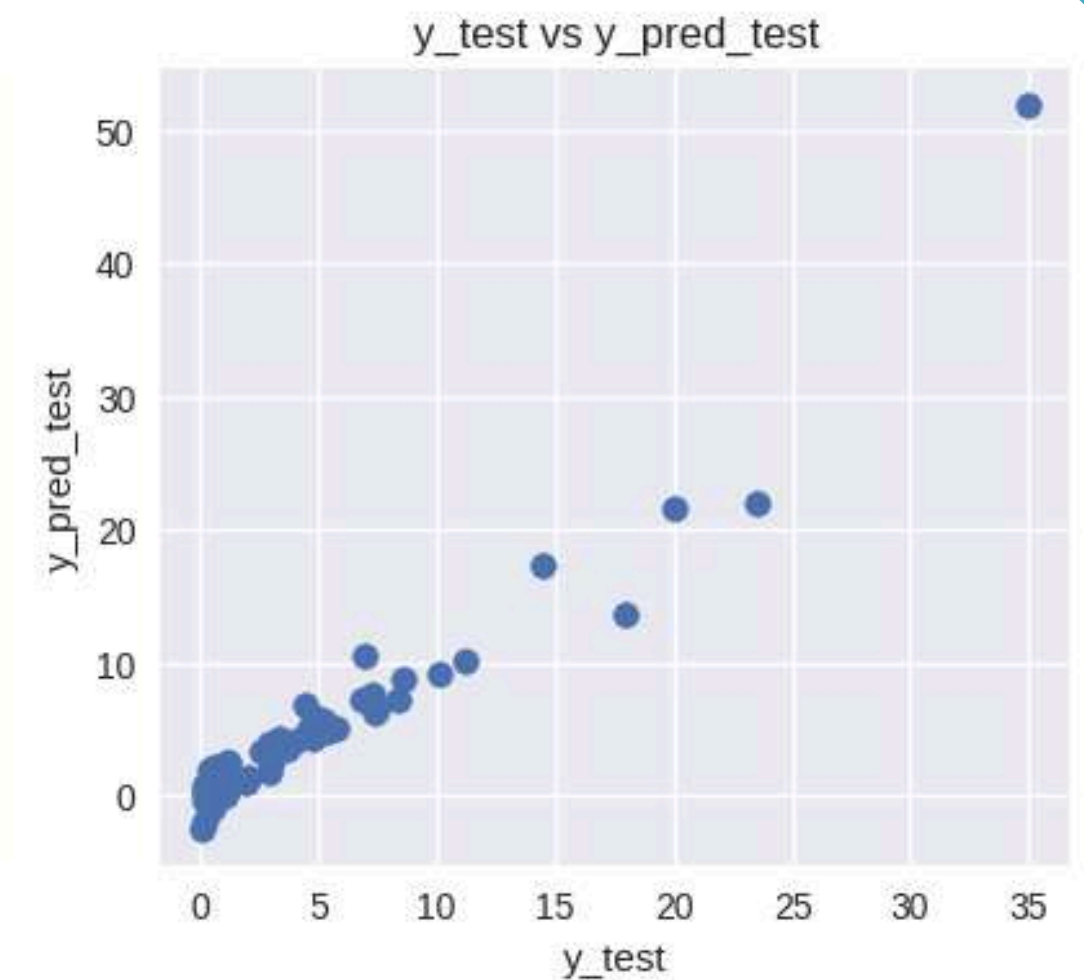
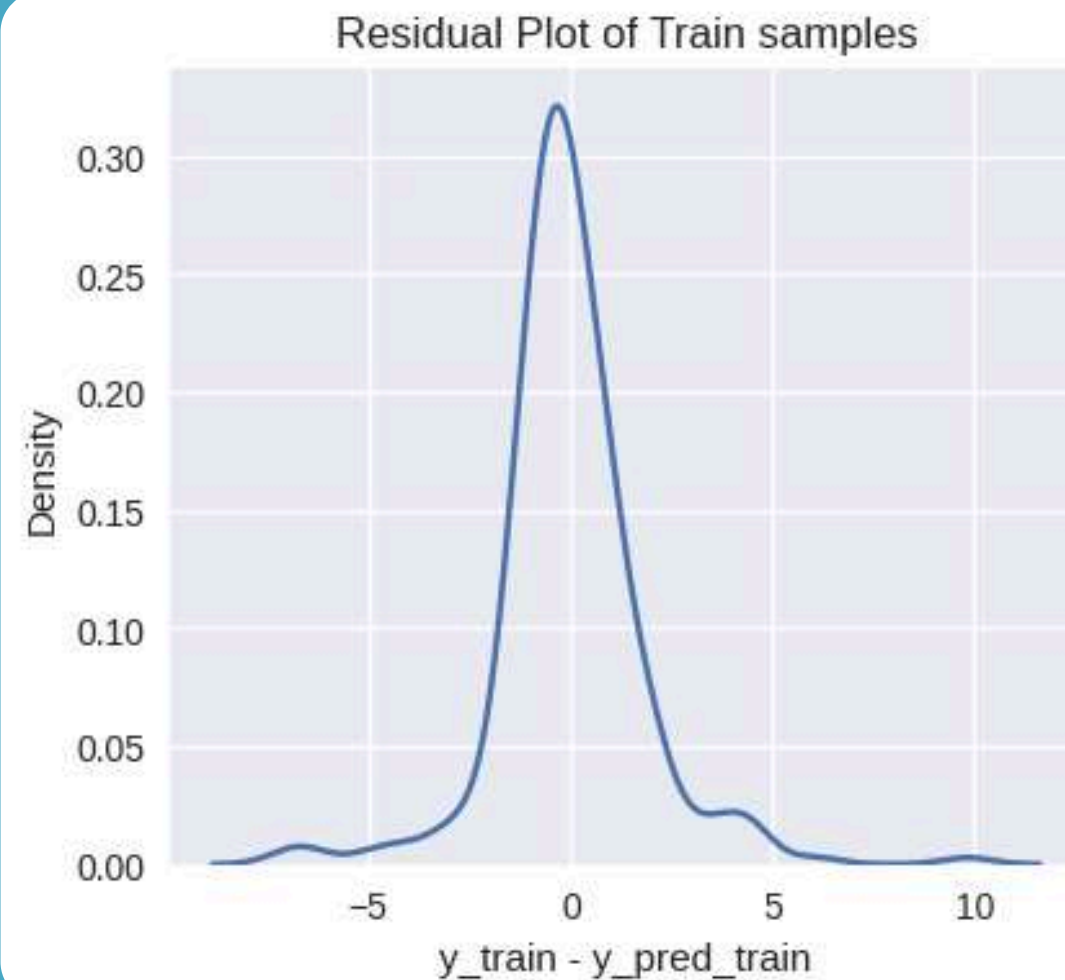
CV mean = 0.81

→ cukup konsisten di berbagai subset data

Train R^2 -score : 0.86
Test R^2 -score : 0.87

Train CV scores :
[0.82140844 0.77218294 0.79144767
0.76785737 0.92113259]
Train CV mean : 0.81

RIDGE

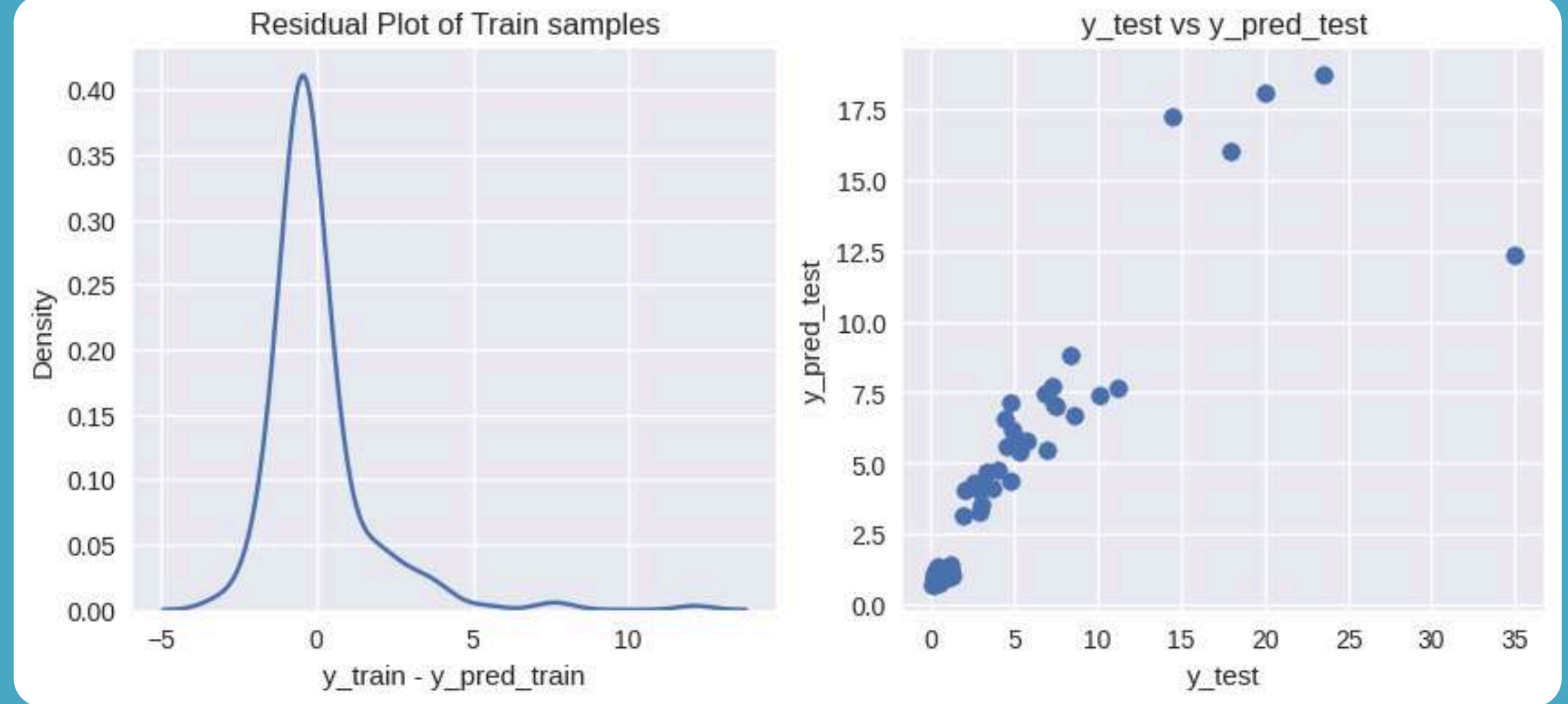


Ridge tidak meningkatkan R^2 secara signifikan, tapi performanya lebih stabil (Train \approx Test)

Train dan Test R^2 seimbang
→ tanda bahwa regularisasi memang menahan model agar tidak terlalu “ngepas” dengan data train

Scatter plot y_test vs y_pred_test:
pola mirip Linear Regression, masih ada outlier harga tinggi (y_test ~35, diprediksi ~50).

RANDOM FOREST



Train $R^2 = 0.96$
→ hampir sempurna di data training.

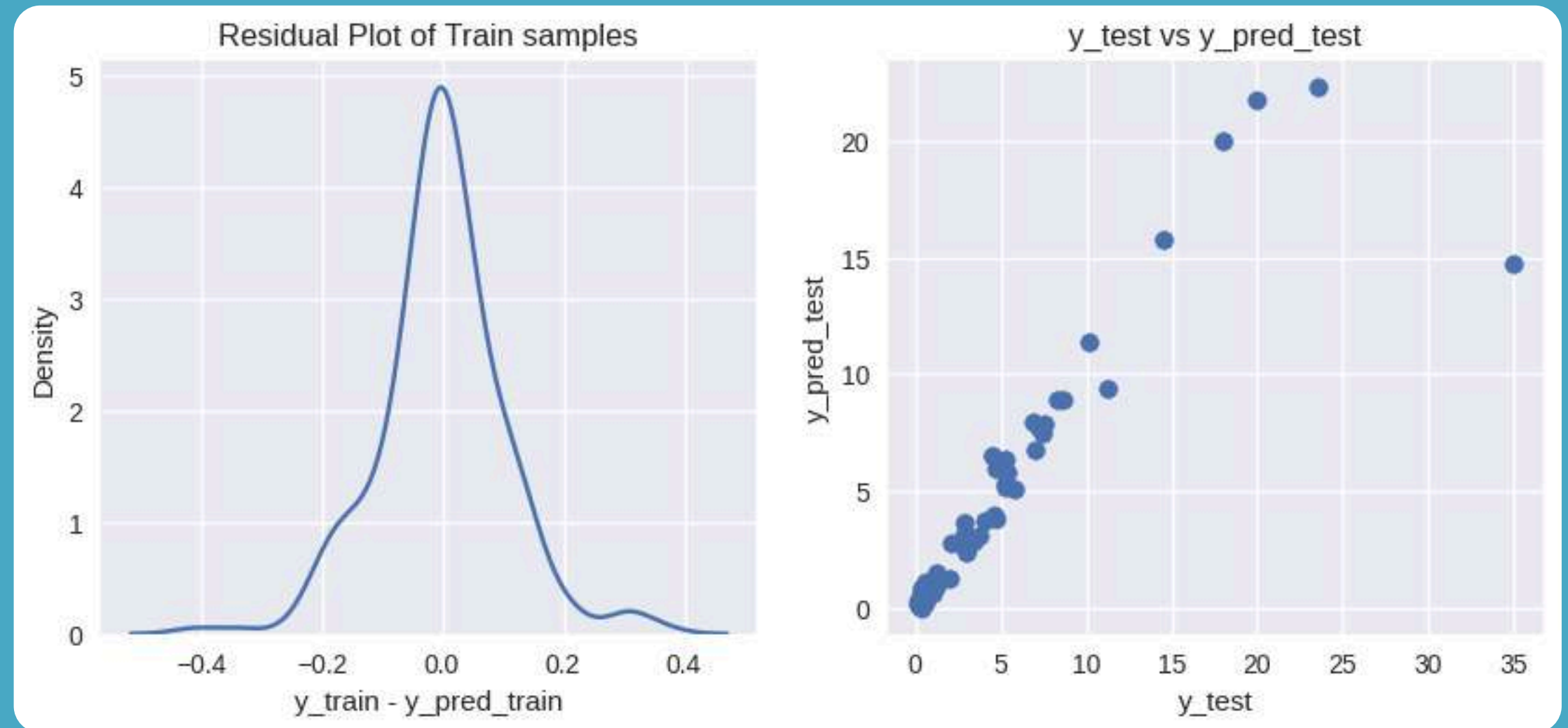
Test performance lebih rendah (0.78)
Selisih besar antara Train (0.96) dan Test (0.78) → indikasi overfitting.

Model belajar terlalu detail dari data training, sehingga generalisasi ke data test berkurang.

Train R^2 -score : 0.87
Test R^2 -score : 0.74

**Train CV scores : [0.85459566
0.8074922 0.92807787 0.85278547
0.90429203]**
Train CV mean : 0.87

GRADIENT BOOSTING



Training sangat sempurna ($R^2 = 1.0$)

Model hampir 100% pas di data training → indikasi overfitting parah


Test R^2 hanya 0.75

- Lebih rendah dari Random Forest (0.78) maupun Linear/Ridge (0.85–0.87)
- Artinya meskipun training fit, generalisasi ke data baru kurang bagus.

**Train R2-score : 1.0
Test R2-score : 0.82**

**Train CV scores : [0.92179717
0.90592312 0.94596514 0.94519004
0.94948663]
Train CV mean : 0.93**

Kesimpulan

	Model	R Squared(Train)	R Squared(Test)	CV score mean(Train)
0	 LinearRegression	0.86	0.87	0.81
1	Ridge	0.85	0.84	0.82
2	RandomForestRegressor	0.87	0.74	0.87
3	GradientBoostingRegressor	1.00	0.82	0.93

Implications

Reliable

Stable predictions across different data
Fast inference = real-time aplikasi

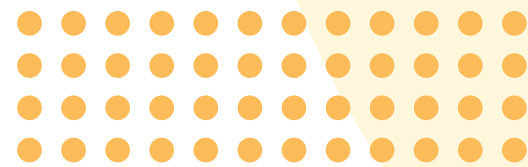
Easy Maintenance

Simple model = easier debugging
Less prone to data drift
Lower computational cost

Gap hampir 0 = perfect generalization

Test > Train = model tidak overfitting sama sekali

Consistent performance across train-test



Rekomendasi

Model tidak hanya membantu dealer dalam menentukan harga beli dan jual yang optimal untuk menjaga margin keuntungan, tetapi juga meningkatkan kepercayaan konsumen melalui informasi harga yang jelas.

integrasi prediksi harga ke dalam platform digital akan memperkuat daya saing bisnis, menarik lebih banyak pengguna, serta mempercepat proses transaksi.

▶ Penentuan Harga Fair

Bantu dealer & individu menetapkan harga optimal sesuai pasar. Dealer dapat memanfaatkan model ini untuk menentukan harga beli mobil dari konsumen (trade-in) agar margin keuntungan tetap sehat.

▶ Meningkatkan Kepercayaan Konsumen

Tampilkan prediksi harga vs harga jual → label Fair / Overpriced / Good Deal.

▶ Segmentasi Pasar

Dealer bisa gunakan insight ini untuk memprioritaskan stok:

- Mobil tahun baru + automatic → margin lebih besar.
- Mobil lama + manual → cocok untuk pasar second-tier (pelajar, pembeli di daerah).

▶ Ekspansi Digital

→ Integrasikan model ke platform jual beli online
→ fitur prediksi harga instan.

THANK
YOU

