



MODEL MACHINE LEARNING

**untuk Estimasi Harga Mobil Bekas
di Arab Saudi pada Syarah.com**

Kirana Azhura

DAFTAR ISI

✓ **Pemahaman Bisnis**

✓ **Pemahaman Data**

✓ **Pembersihan Data**

✓ **Analisis Data Eksplorasi**

Pra-Pemrosesan Data ✓

Pemodelan ✓

Interpretasi Model ✓

Kesimpulan dan Saran ✓



PEMAHAMAN BISNIS



PEMAHAMAN BISNIS

Syarah.com adalah platform e-commerce otomotif terkemuka di Arab Saudi yang menyediakan layanan jual beli mobil baru dan bekas secara online. Didirikan pada tahun 2015 dan berbasis di Riyadh, Syarah telah merevolusi cara konsumen membeli mobil dengan menawarkan pengalaman yang mudah, aman, dan terpercaya.



PEMAHAMAN BISNIS

Latar Belakang

- Pasar mobil bekas di Arab Saudi terus tumbuh karena tingginya permintaan akan kendaraan terjangkau.
- Penentuan harga yang tepat sangat menantang karena banyaknya faktor yang memengaruhi.
- **Syarah.com**, sebagai platform jual beli mobil, ingin meningkatkan layanan dengan menyediakan estimasi harga berbasis teknologi.



MASALAH

- Harga mobil bekas sangat bervariasi dan sering ditentukan secara subyektif.
- Tidak adanya sistem penilaian harga otomatis menyebabkan risiko overprice atau underprice.
- Evaluasi manual tidak efisien bagi platform yang menangani ribuan mobil setiap hari.

TUJUAN ANALISIS

- Membangun model prediksi harga mobil bekas berbasis machine learning.
- Mengidentifikasi faktor-faktor kunci yang memengaruhi harga (tahun, kilometer, merk, dll).
- Menyediakan sistem estimasi otomatis untuk:
 - Penjual: Menentukan harga yang kompetitif
 - Pembeli: Mendapatkan acuan harga wajar
 - Perusahaan: Meningkatkan efisiensi & profitabilitas

STAKEHOLDER

Tim Produk dan Teknologi Syarah.com

➤ Tim Produk dan Teknologi merupakan pihak internal yang bertanggung jawab atas pengembangan, implementasi, dan pemeliharaan fitur-fitur baru di platform Syarah.com, termasuk integrasi sistem machine learning untuk estimasi harga mobil bekas.

EVALUASI METRIK

- MAE (Mean Absolute Error):
Mengukur rata-rata selisih absolut antara prediksi dan nilai aktual (dalam riyal).
➤ Mudah dipahami & robust terhadap outlier.
- MAPE (Mean Absolute Percentage Error):
Mengukur rata-rata kesalahan dalam bentuk persentase.
➤ Memudahkan perbandingan performa antar model & harga mobil.
- Kesimpulan:
Kombinasi MAE + MAPE memberikan evaluasi akurat dan mudah dipahami, baik dari sisi nominal maupun persentase kesalahan.



PEMAHAMAN DATA



»» PEMAHAMAN DATA ««

- Dataset ini terdiri dari 5624 data mobil bekas yang dikumpulkan dari situs syarah.com.
- Setiap baris dalam dataset mewakili satu unit mobil bekas.
- Terdapat 10 kolom yang menjelaskan informasi mobil bekas.



PEMAHAMAN DATA



Kolom	Deskripsi
Type	Jenis mobil (sedan, SUV, dll)
Region	Lokasi penjualan di Arab Saudi
Make	Merek mobil (Toyota, Hyundai, dll)
Gear_Type	Tipe transmisi (otomatis/manual)
Origin	Asal mobil (lokal atau impor)
Options	Fitur tambahan (ABS, Airbag, dll)
Year	Tahun pembuatan mobil
Engine_Size	Kapasitas mesin (liter/cc)
Mileage	Jarak tempuh mobil (dalam kilometer)
Negotiable	Harga bisa dinego (`True` jika Price = 0)
Price	Harga mobil bekas (SAR – Saudi Riyal)



PEMBERSIHAN DATA



PEMBERSIHAN DATA



Data mobil bekas Syara.com memiliki 10 kolom dengan 6 kolom object dan 4 numerik, yang berisikan 5624 baris dan tidak memiliki nilai kosong.



Terdapat 3 data duplikat. Penghapusan data duplikat dilakukan untuk mencegah bias, menghindari overfitting, dan memastikan model mempelajari informasi yang unik dan relevan sehingga prediksi harga mobil bekas menjadi lebih akurat.



Terdapat 31.9% data dengan harga = 0 yang seluruhnya memiliki status Negotiable = True, sedangkan semua data dengan harga > 0 memiliki Negotiable = False, membentuk perfect separation. Kondisi ini dapat menyebabkan distorsi pada model, seperti koefisien ekstrem dan error statistik yang tidak masuk akal. Oleh karena itu, baris dengan Price = 0 dan kolom Negotiable dihapus untuk menjaga kualitas dan stabilitas model.



»» PEMBERSIHAN DATA ««

- Drop Harga mobil < 10000 SAR
- Drop Origin = Unknown

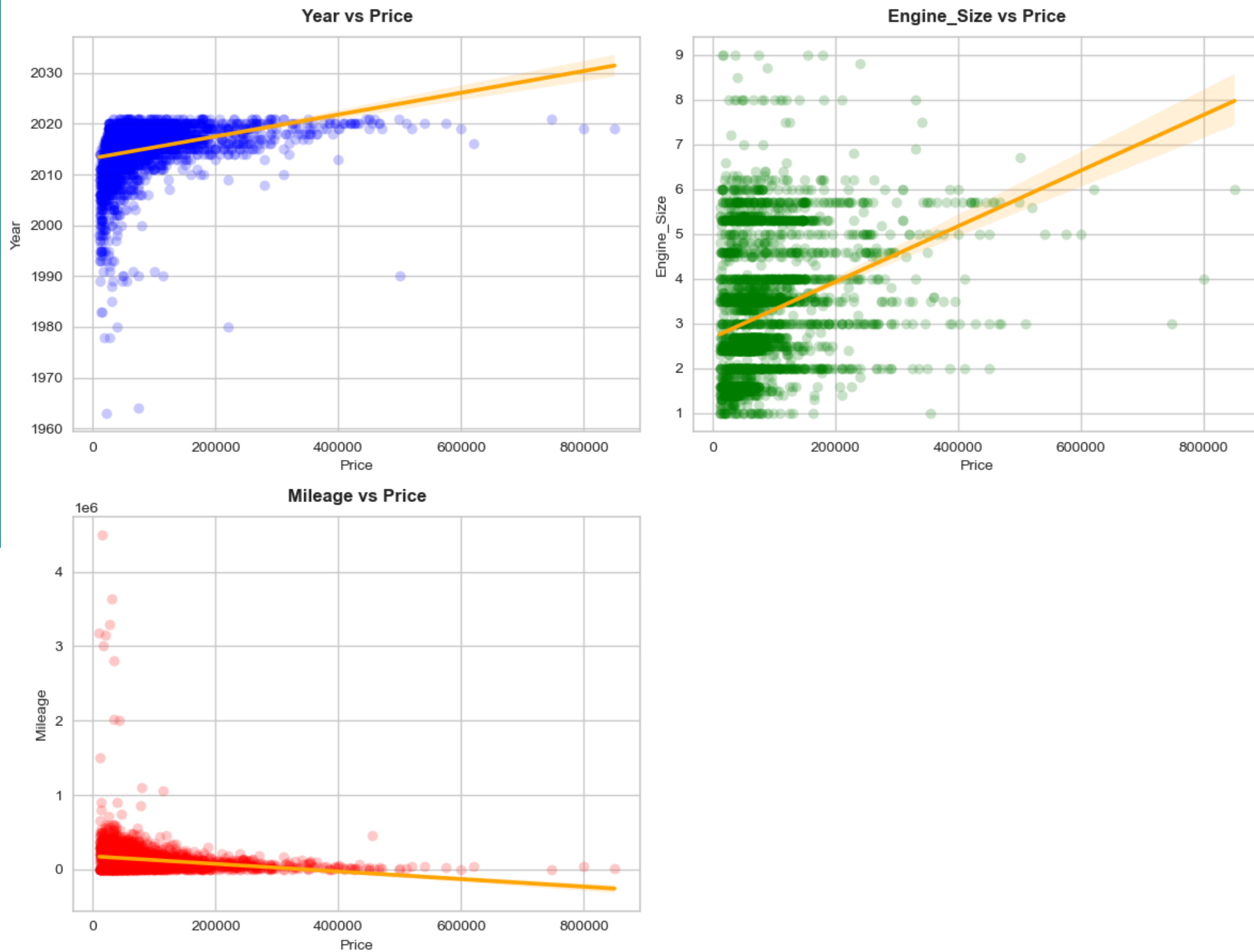


ANALISIS DATA EKSPLORASI



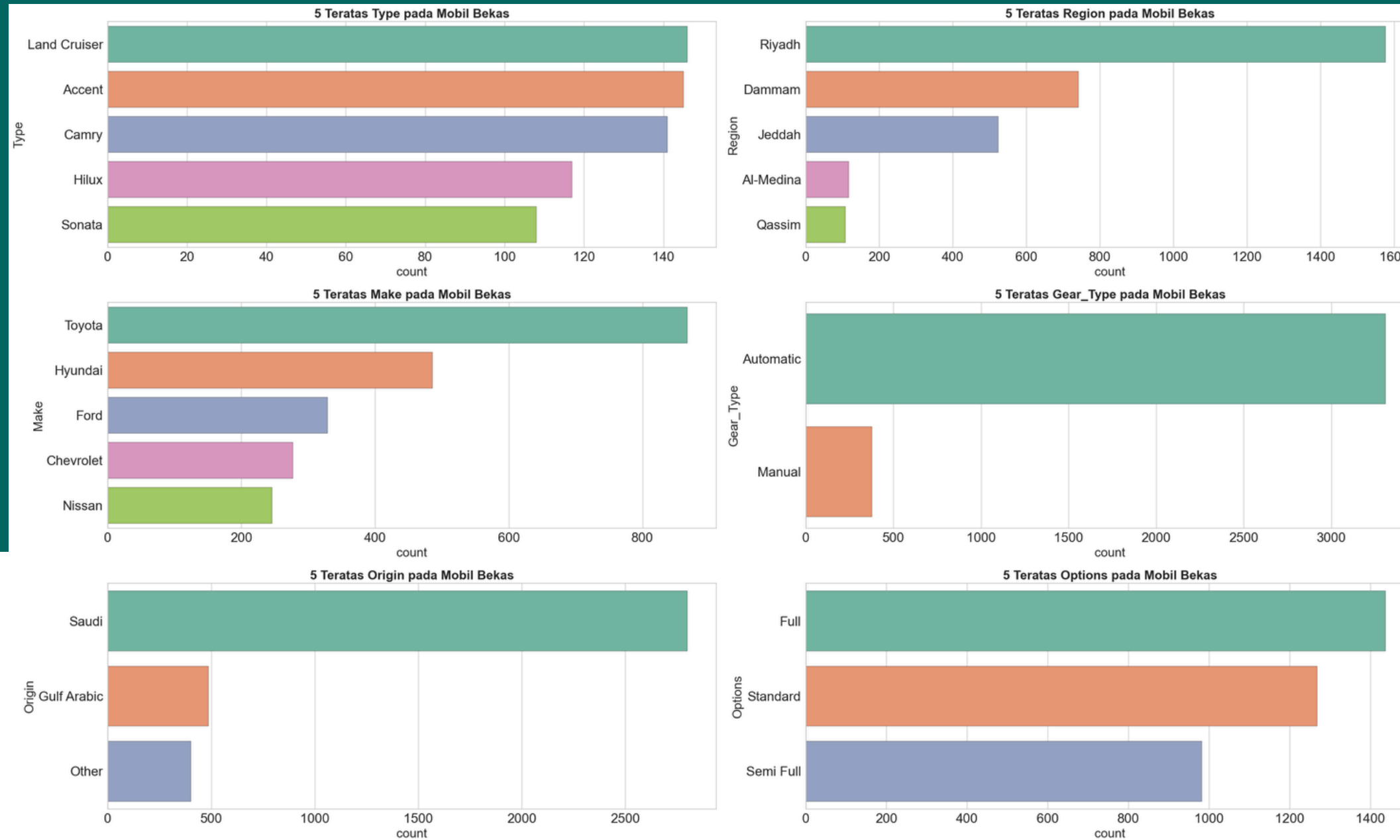
ANALISIS DATA EKSPLORASI

Scatter Plots Harga dengan Garis Regresi



- Tahun Produksi vs Harga:
 - > Semakin baru tahun mobil, semakin tinggi harganya.
- Ukuran Mesin vs Harga:
 - > Mesin lebih besar → harga lebih mahal (umumnya mobil premium/SUV).
- Mileage vs Harga:
 - > Semakin jauh jarak tempuh, harga makin turun.

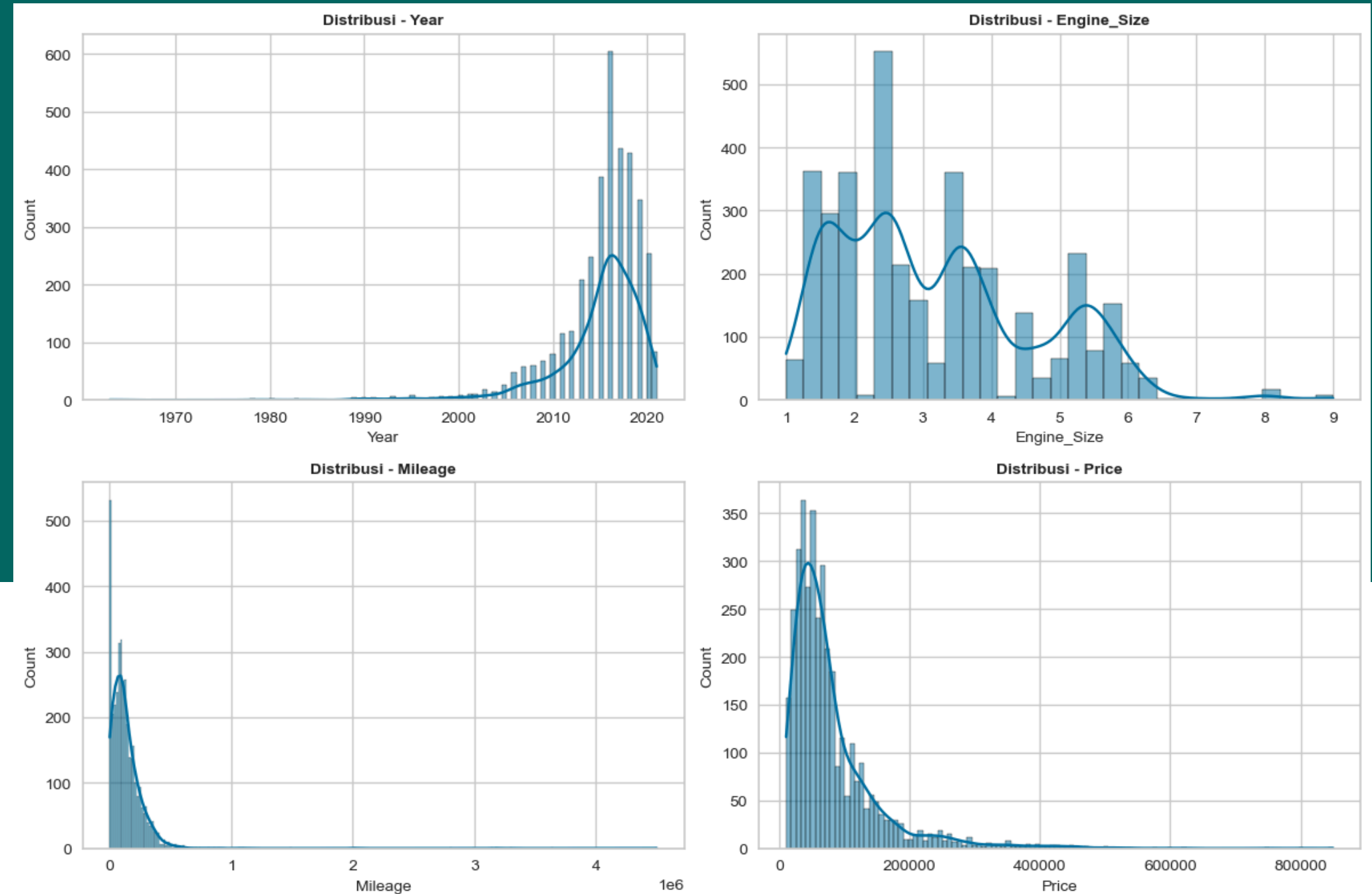
ANALISIS DATA EKSPLORASI



- Tipe Terpopuler: Land Cruiser, Accent, Camry mendominasi — SUV dan sedan paling banyak dijual.
- Wilayah Terbanyak: Riyadh, Dammam, dan Jeddah jadi pusat pasar mobil bekas.
- Merek Favorit: Toyota, Hyundai, dan Ford unggul — merek Jepang dan Korea dipercaya.
- Transmisi: Automatic jauh lebih banyak dari Manual.
- Asal Mobil: Mayoritas dari Saudi, disusul Gulf Arabic dan Other.
- Opsi Fitur: Full dan Standard paling umum, menunjukkan pentingnya fitur tambahan.

ANALISIS DATA EKSPLORASI

- Year:
Mayoritas mobil adalah keluaran baru (setelah 2010), fokus pada kendaraan modern.
- Engine Size:
Ukuran mesin kecil-menengah (1.0–2.5L) mendominasi; mesin besar sangat jarang.
- Mileage:
Mobil dengan jarak tempuh rendah paling banyak; mileage tinggi adalah outlier.
- Price:
Harga didominasi di bawah 200.000; hanya sedikit mobil mewah berharga tinggi.



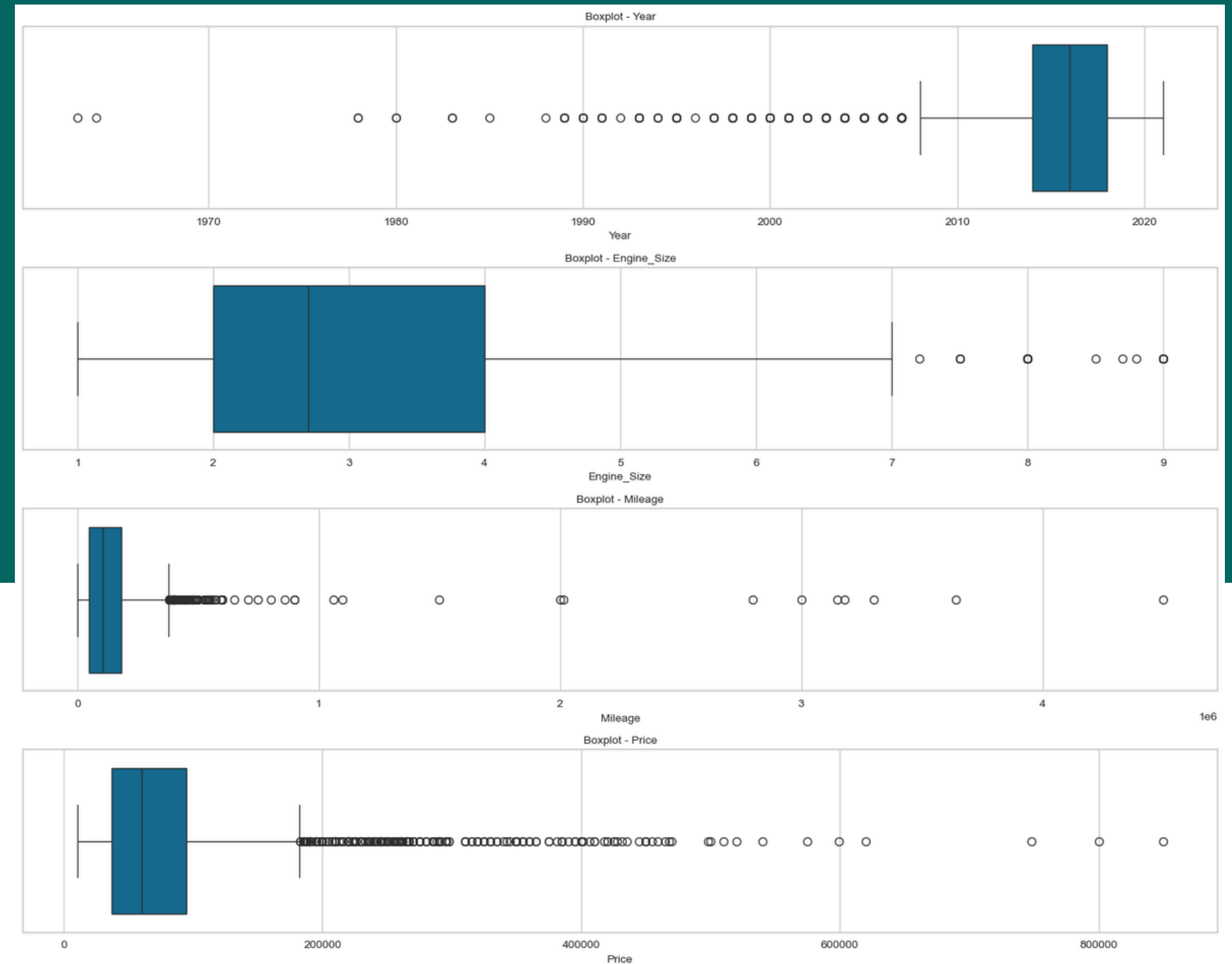
ANALISIS DATA EKSPLORASI

- Year:

Mobil dengan tahun produksi sebelum 2000 dihapus dari analisis karena tidak relevan dengan pasar mobil bekas modern, berpotensi mendistorsi statistik, dan memiliki perbedaan teknis signifikan

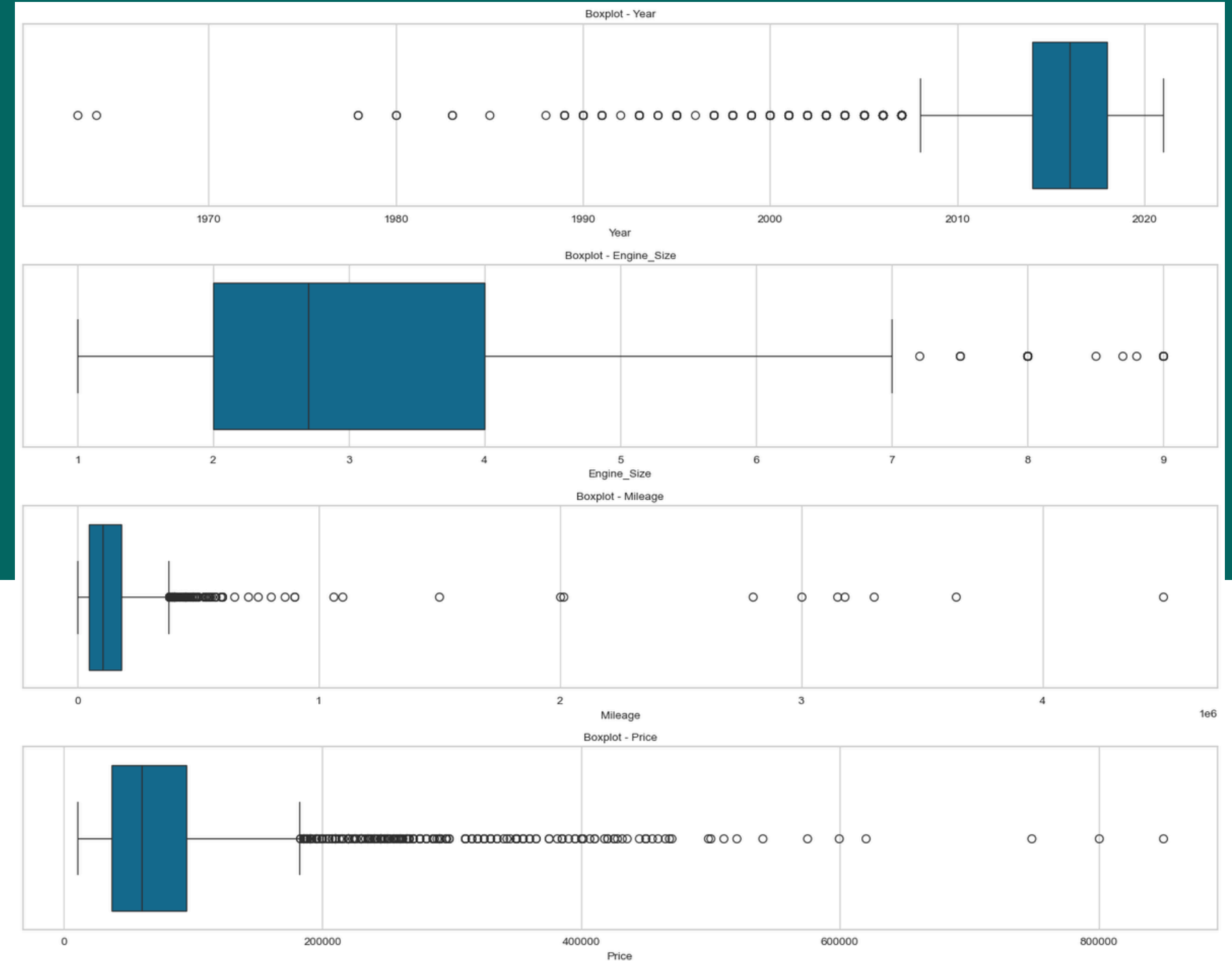
- Engine Size:

Data dengan Engine Size > 7.0L dihapus karena kemungkinan besar merupakan hasil kesalahan input, tidak realistis untuk mobil penumpang umum, dan berisiko menjadi outlier yang mengganggu validitas analisis.

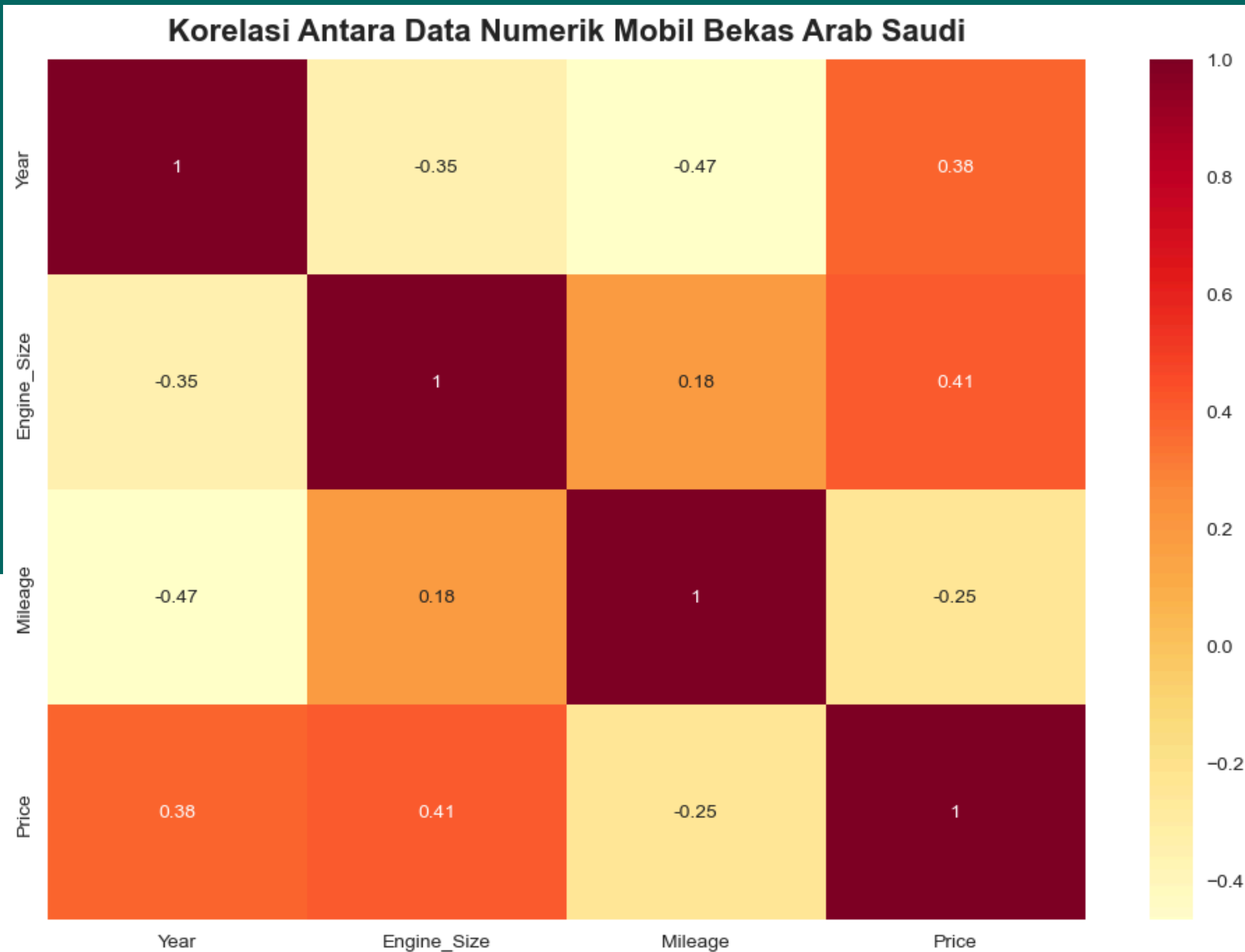


ANALISIS DATA EKSPLORASI

- Mileage:
Mileage di atas 379.000 km dihapus karena dianggap outlier yang tidak merepresentasikan kondisi pasar mobil bekas, berpotensi kesalahan input, dan dapat menurunkan akurasi analisis serta model prediktif.
- Price:
Harga mobil di atas 269.000 SAR dihapus karena dianggap outlier ekstrem yang dapat mengganggu akurasi model, distorsi statistik, dan tidak merepresentasikan pasar mobil bekas secara umum.



ANALISIS DATA EKSPLORASI



- Year vs. Price
Korelasi positif sedang ($r = 0.38$) — mobil lebih baru cenderung lebih mahal.
- Engine Size vs. Price
Korelasi positif moderat ($r = 0.41$) — mesin besar biasanya punya harga lebih tinggi.
- Mileage vs. Price
Korelasi negatif lemah ($r = -0.25$) — jarak tempuh tinggi menurunkan nilai mobil.



PRA-PEMROSESAN DATA



PRA-PEMROSESAN DATA

1

Menentukan X dan y

y : target > harga mobil

X : fitur-fitur yang dianggap memengaruhi harga mobil (9 kolom lainnya)

2

Membagi data menjadi data Train dan data Test dengan proporsi 70 : 30

3

Melakukan Encoding:

Binary Encoding: Type, Region, Make

Onehot Encoding: Gear_Type, Origin, Options

4

Melakukan Scaling:

Robust Scaling: Year, Engine_Size, Mileage



PEMODELAN

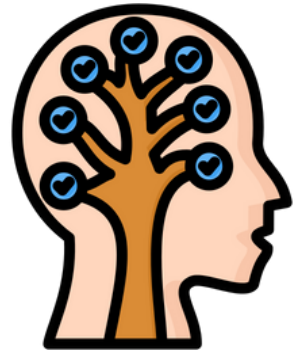


»» BENCHMARK MODEL ««

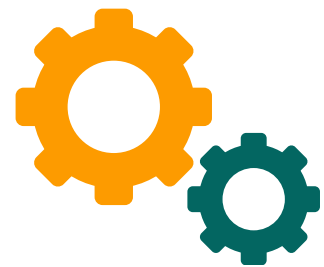
	model	mean_mae	std_mae	mean_mape	std mape
5	XGBRegressor(base_score=None, booster=None, ca...	-14195.436089	376.887644	-0.245162	0.003734
7	RandomForestRegressor(random_state=42)	-15006.704073	697.036814	-0.262003	0.009221
6	GradientBoostingRegressor(random_state=42)	-16662.726007	833.444407	-0.282025	0.015361
3	KNeighborsRegressor()	-16277.010753	447.023226	-0.302071	0.009482
4	DecisionTreeRegressor(random_state=42)	-19296.379716	1211.037028	-0.323033	0.009979
1	Ridge(random_state=42)	-23572.880845	891.902089	-0.443289	0.021717
2	Lasso(random_state=42)	-23579.062098	892.509293	-0.443512	0.021754
0	LinearRegression()	-23580.661709	892.822271	-0.443561	0.021750

XGBoost dipilih sebagai model terbaik karena memiliki nilai RMSE, MAE, dan MAPE terendah serta deviasi standar yang kecil, menunjukkan akurasi tinggi dan kinerja yang stabil.

MODEL TERBAIK



XGBoost (Extreme Gradient Boosting) adalah algoritma machine learning berbasis tree boosting yang dibangun secara bertahap untuk meminimalkan kesalahan prediksi. Setiap pohon baru memperbaiki kelemahan pohon sebelumnya.



Karakteristik utamanya meliputi kemampuan menangani data non-linear, efisiensi tinggi dalam komputasi, dukungan terhadap regularisasi (L1 dan L2) untuk mencegah overfitting, serta mampu mengatasi missing value secara otomatis. Model ini sangat cocok untuk prediksi dengan data kompleks dan menghasilkan performa yang stabil dan akurat.



XGBoost unggul karena akurasi tinggi, mampu menangani data non-linear, memiliki mekanisme regularisasi untuk mencegah overfitting, serta efisien pada dataset besar.

HYPERPARAMETER TUNING

PARAMETER TUNING


```
hyperparam = {  
    'model__max_depth': [2, 3, 4],  
    'model__learning_rate': [0.01, 0.02, 0.03],  
    'model__n_estimators': [100, 200, 300],  
    'model__subsample': [0.5, 0.6, 0.7],  
    'model__colsample_bytree': [0.5, 0.6],  
    'model__gamma': [5, 10],  
    'model__reg_alpha': [5, 10],  
    'model__reg_lambda': [5, 10],  
    'model__min_child_weight': [10, 20]  
}
```




PARAMETER TERBAIK

```
{'model__colsample_bytree': 0.6,  
 'model__gamma': 5,  
 'model__learning_rate': 0.03,  
 'model__max_depth': 4,  
 'model__min_child_weight': 10,  
 'model__n_estimators': 300,  
 'model__reg_alpha': 10,  
 'model__reg_lambda': 5,  
 'model__subsample': 0.5}
```

PERBANDINGAN ERROR MODEL



	Model	MAE Test	MAE Train	MAPE Test	MAPE Train
0	XGBoost after tuning	16731.115969	13166.393411	0.248198	0.214555
0	XGBoost before tuning	14892.361914	2991.940954	0.232836	0.049580



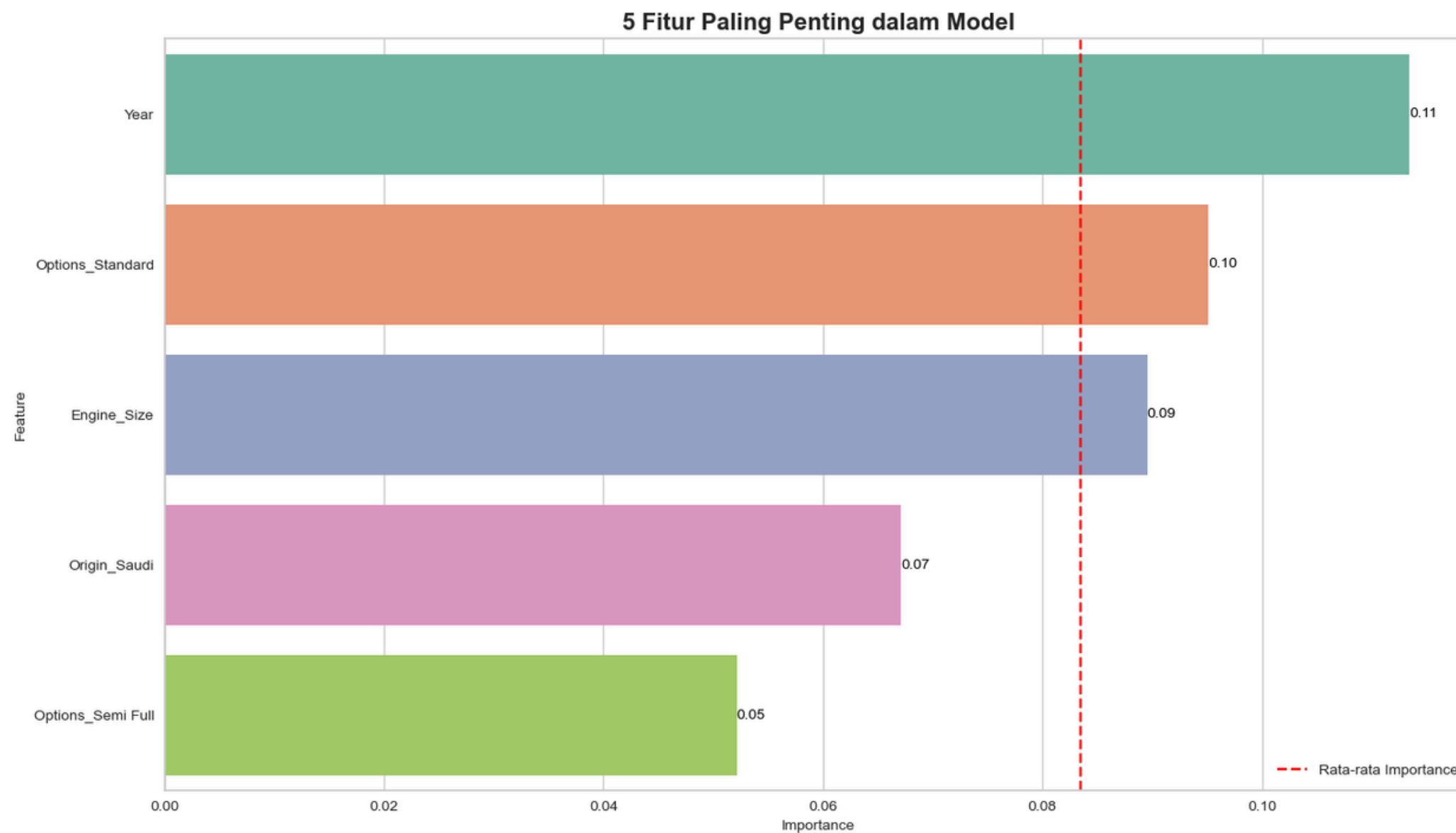
Setelah dilakukan hyperparameter tuning, model XGBoost menunjukkan penurunan performa pada data testing namun menghasilkan generalisasi yang lebih baik. Hal ini terlihat dari peningkatan MAE dan MAPE pada training set yang signifikan, menandakan berkurangnya overfitting. Meskipun terjadi sedikit kenaikan MAE dan MAPE pada test set, nilainya tetap dalam kisaran wajar dan lebih seimbang dengan hasil training. Trade-off ini menunjukkan bahwa tuning berhasil menciptakan model yang lebih realistis dan andal saat dihadapkan pada data baru.



INTERPRETASI MODEL

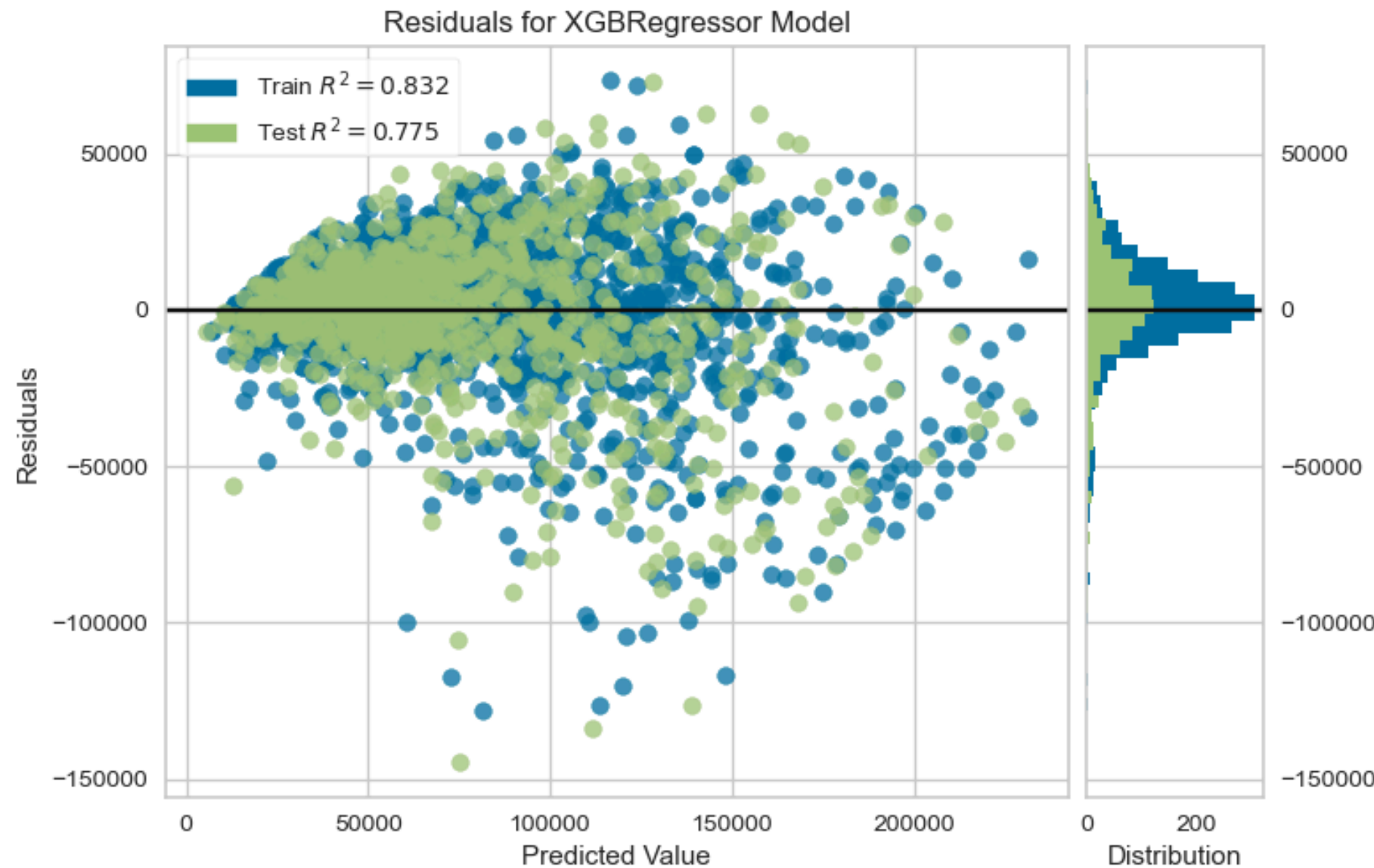


FEATURE IMPORTANCE



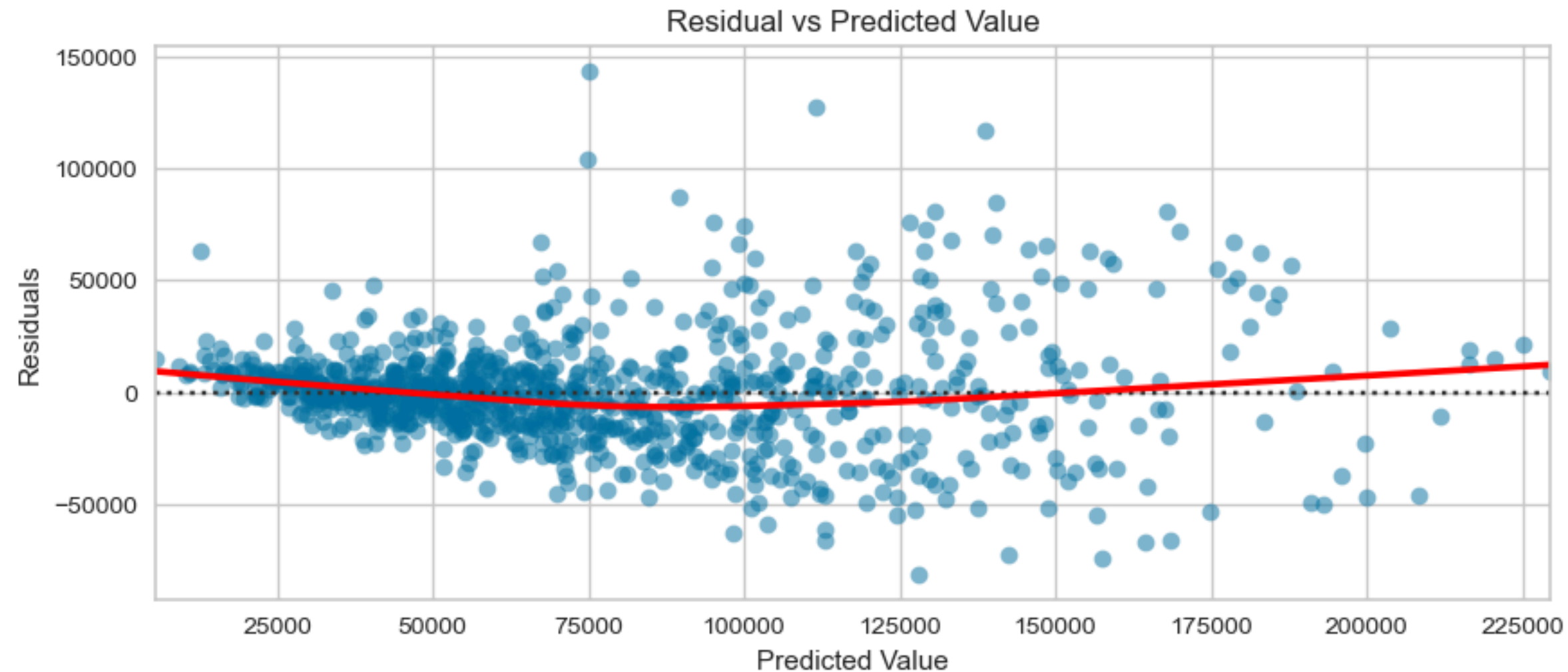
- Year (0.11)
→ Mobil yang lebih baru cenderung lebih mahal.
- Options_Standard (0.10)
→ Fitur standar menaikkan nilai jual kendaraan.
- Engine_Size (0.09)
→ Mesin lebih besar → harga lebih tinggi.
- Origin_Saudi (0.07)
→ Mobil asal Saudi lebih bernilai di pasar lokal.
- Options_Semi Full (0.05)
→ Opsi tambahan memberi nilai tambah, meski tidak sebesar opsi standar.

RESIDUAL PLOT



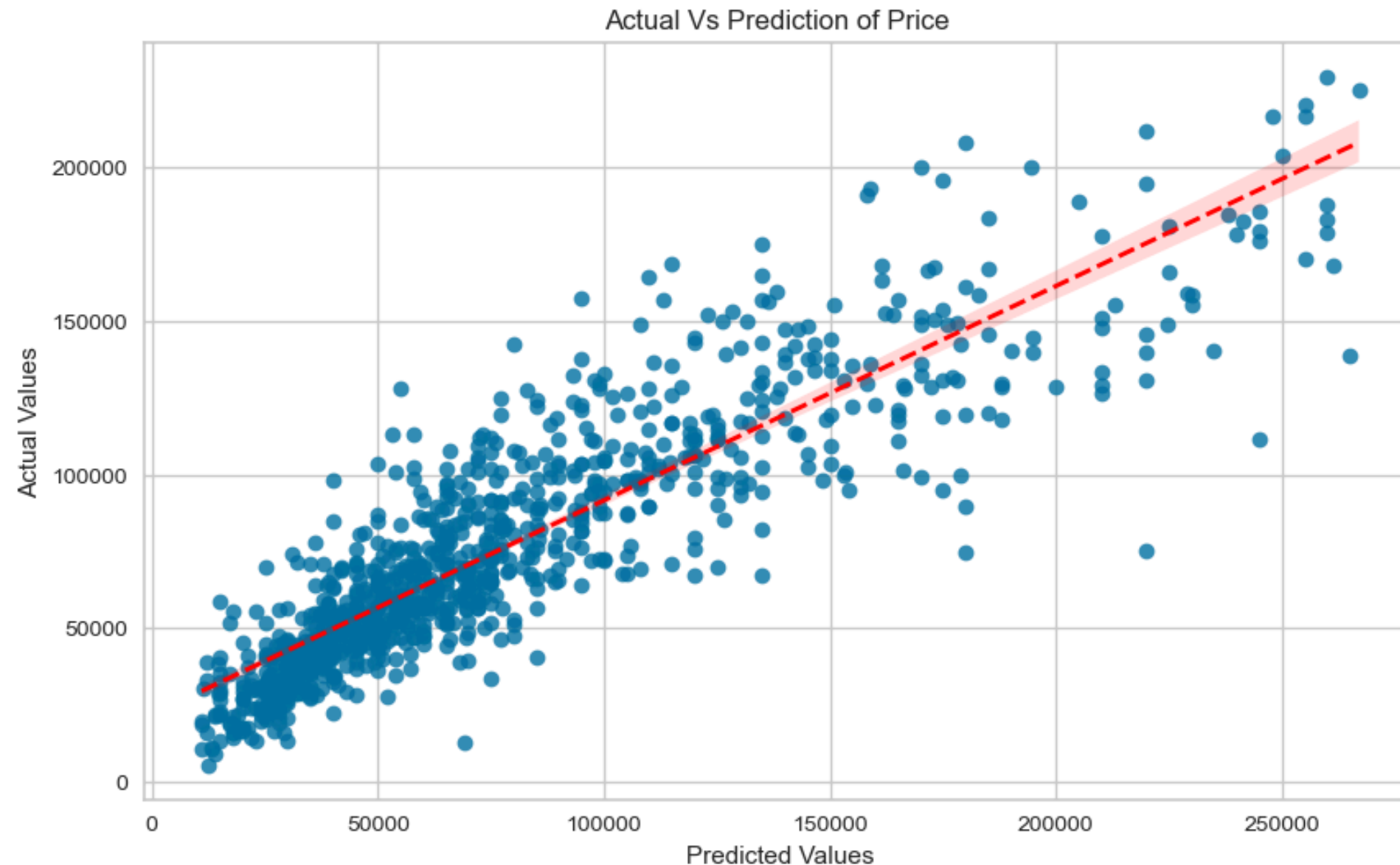
- Residual model XGBoost tersebar cukup simetris di sekitar nol, menunjukkan prediksi yang cukup akurat.
- Nilai R^2 pada training (0.832) dan testing (0.775) menunjukkan model memiliki generalisasi yang baik tanpa overfitting signifikan.
- Sebagian besar error berada dalam rentang ± 50.000 , dengan outlier pada mobil berharga tinggi, mengindikasikan potensi keterbatasan model di segmen premium.

RESIDUAL VS PREDICTED VALUE PLOT



- Pola Lengkung menunjukkan hubungan non-linear yang belum sepenuhnya ditangkap model.
- Heteroskedastisitas: Error meningkat di harga tinggi → indikasi model kurang stabil di segmen mahal.
- Underestimate pada mobil mahal (residual negatif dominan).
- Prediksi Akurat di harga rendah-menengah (residual kecil & seimbang).

ACTUAL VS PREDICTION PLOT

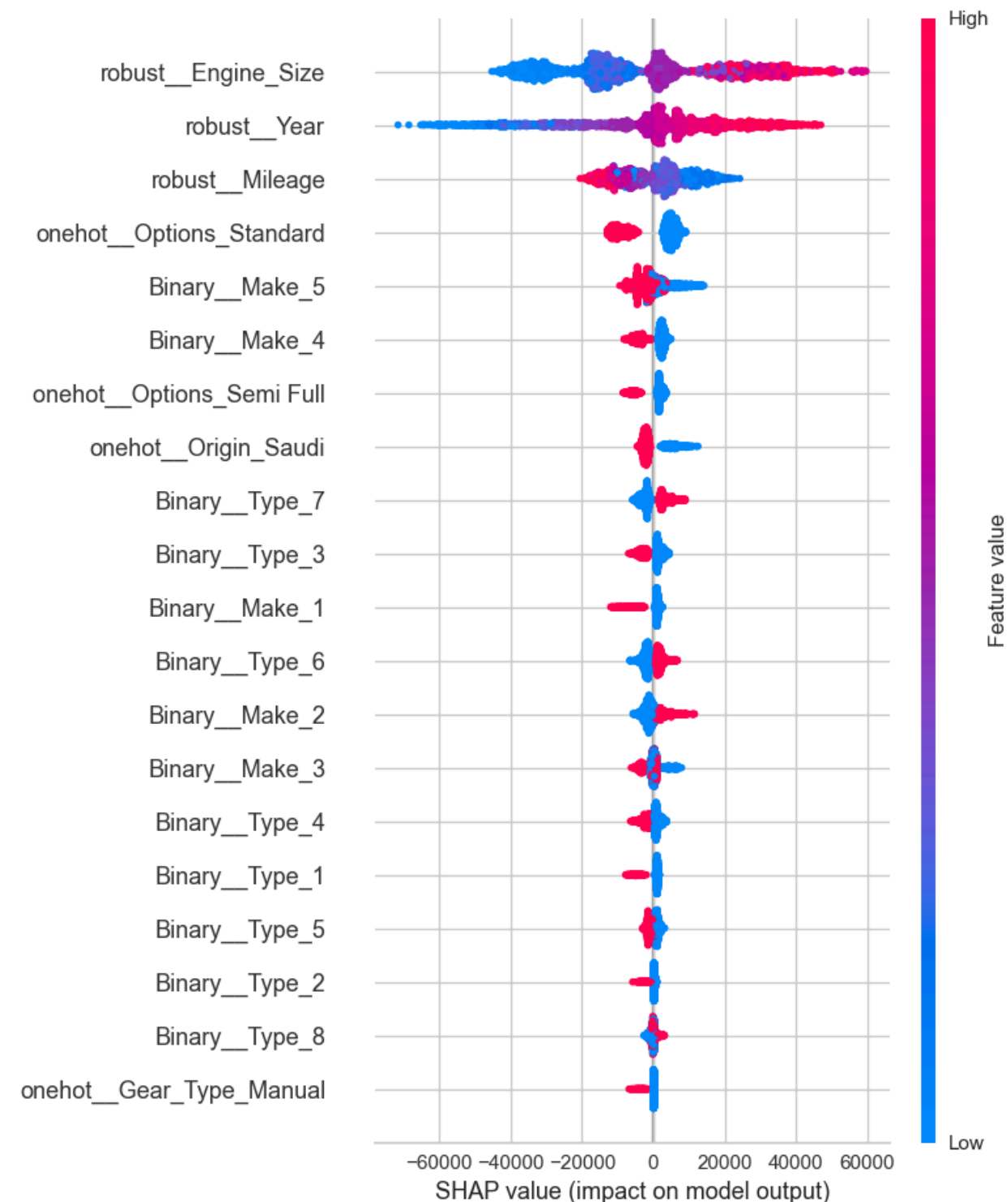


Model menunjukkan hubungan linear yang kuat, dengan titik-titik prediksi tersebar dekat garis 45° — menandakan akurasi tinggi terutama pada segmen harga menengah (25.000–150.000).

Namun, untuk harga mobil >150.000 , model mulai underpredict, mengindikasikan kesulitan dalam memodelkan mobil mewah.

Secara keseluruhan, tidak terdapat outlier sistematis yang mencolok, mencerminkan data telah dibersihkan dengan baik.

INTERPRETASI MODEL



- Fitur **Engine Size, Year, dan Mileage** adalah tiga faktor terpenting dalam menentukan harga mobil bekas.
- Fitur-fitur kategorikal seperti **merek, tipe, opsi, dan asal** memiliki pengaruh sekunder.
- Fitur dengan nilai tinggi bisa berdampak positif atau negatif tergantung konteks (misalnya Mileage tinggi = buruk, tapi Engine Size tinggi = bagus).
- Grafik ini membuktikan bahwa model tidak hanya bergantung pada satu fitur, tapi menggunakan berbagai dimensi untuk prediksi yang lebih akurat.

PERHITUNGAN BISNIS

Parameter	Nilai Asumsi
Jumlah mobil masuk ke platform/bulan	5.000 unit
Persentase mobil dengan harga tidak wajar (tanpa model)	10% (500 mobil)
Rata-rata kerugian karena mispricing	±SAR 10.000 per mobil
Akurasi model prediksi harga	Meningkatkan validasi ±75.2% akurat
Biaya operasional model (server + tenaga ahli dsb)	SAR 25.000 per bulan

Skenario	Kerugian Mispricing	Biaya Operasional	Total Cost
Tanpa model	SAR 5.000.000	SAR 0	SAR 5.000.000
Dengan model XGBoost	SAR 1.240.000	SAR 25.000	SAR 1.265.000

Dengan menggunakan model prediksi harga:

- Syarah.com dapat menghemat sekitar SAR 3.735.000 per bulan.

- Selain efisiensi biaya, keuntungan lain:
 - + Meningkatkan kepercayaan pengguna.
 - + Menekan penipuan harga dan manipulasi.
 - + Membantu standarisasi pasar mobil bekas.



**KESIMPULAN,
SARAN, LIMITASI**



KESIMPULAN

MODEL

Model XGBoost berhasil dibangun untuk memprediksi harga mobil bekas dengan performa yang cukup baik. MAE Test ~16.731 SAR dan MAPE Test ~24.8% menunjukkan akurasi prediksi yang dapat diterima di pasar mobil bekas. Model cukup stabil (tidak overfit).

FITUR-FITUR BERPENGARUH

Harga mobil bekas paling dipengaruhi oleh tahun produksi, kelengkapan mobil, ukuran mesin, asal mobil, dan kilometer tempuh.

BISNIS

Penggunaan model prediksi membantu menekan deviasi harga dari $\pm 15\%$ menjadi 8–10%, meningkatkan akurasi estimasi, efisiensi biaya, dan potensi profit, serta mengurangi risiko overprice dan underprice dalam jual beli mobil bekas.

REKOMENDASI DAN LIMITASI

REKOMENDASI

- Gunakan model sebagai alat estimasi harga otomatis di platform Syarah.com.
- Terapkan validasi harga otomatis dan beri notifikasi jika harga menyimpang dari estimasi.
- Perbarui model secara berkala dengan data terbaru agar tetap akurat.
- Tambahkan fitur seperti lokasi, riwayat servis, bahan bakar, dan popularitas untuk meningkatkan prediktivitas.

LIMITASI MODEL

- Belum mempertimbangkan kondisi visual mobil secara langsung.
- Penghapusan outlier bisa menghilangkan mobil mewah yang valid.
- Model cenderung underestimate harga mobil mewah.
- Tidak mempertimbangkan faktor eksternal (musim, promo, ekonomi).
- Data bersifat snapshot, belum mendukung prediksi harga real-time.



TERIMA KASIH

Kirana Azhura