PROJECT MATA KULIAH DATA WRANGLING

ANALISIS HARGA MOBIL BEKAS

DI WILAYAH DKI JAKARTA

Disusun Oleh (Kelompok 12):

Fawwaz Azri Manshur Prasetya	[24031554074]
Meisya Natasafira	[24031554174]

Dosen Pembimbing:

Dinda Galuh Guminta, S.Stat., M.Stat.	[0011129602]
Belgis Ainatul Iza, S.Si., M.Mat.	[202509237]

PROGRAM STUDI S1 SAINS DATA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SURABAYA

2025

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pasar harga mobil bekas di daerah DKI Jakarta berkembang pesat seiring dengan meningkatnya kebutuhan masyarakat terhadap sarana transportasi dengan harga yang terjangkau. Mobil bekas memiliki peran penting dalam pasar otomotif, sehingga memperkirakan harganya menjadi hal yang diperlukan bagi pembeli dan penjual untuk mengetahui nilai yang tepat [3]. Sebagai pusat aktivitas ekonomi dan mobilitas tertinggi di Indonesia, DKI Jakarta menjadi wilayah dengan transaksi mobil bekas dengan harga yang paling beragam. Dengan adanya berbagai platform online penjualan mobil bekas, masyarakat dimudahkan dalam memperoleh informasi harga dengan sangat mudah.

Harga mobil bekas di berbagai platform penjualan seperti carmudi dan caroline menunjukkan variasi yang sangat besar. Harga mobil bekas dipengaruhi oleh faktor seperti brand, tipe, tahun produksi, dan kondisi mobil [1]. Selain itu terdapat data set dari kaggle yang memiliki banyak faktor dalam mempengaruhi harga jual mobil seperti brand, tipe, model, warna, tahun produksi, harga kredit, harga cash, dan masih banyak faktor-faktor yang lain. Perbedaan harga yang signifikan meskipun dalam model yang sama, membuat konsumen kesulitan dalam menentukan harga pasar yang wajar.

Dalam kondisi ini, analisis data menjadi peran penting bagi konsumen, dengan adanya analisis perbandingan harga dapat membantu menilai harga yang sesuai dan membuat keputusan yang tepat. Selain itu data analisis perbandingan harga juga dapat digunakan penjual untuk menentukan harga yang sesuai. Pemilihan DKI Jakarta sebagai lokasi analisis sangat relevan karena tingginya aktivitas penjualan, keragaman jenis mobil, dan ketersediaan data.

Data dari berbagai sumber masih memiliki format yang berbeda, mengandung duplikasi, dan informasi yang kurang lengkap. Salah satu cara untuk memprediksi harga mobil adalah menggunakan machine learning, yaitu melatih model dengan algoritma supervised learning agar mampu menganalisis data dan membuat prediksi secara akurat [2]. Dengan melakukan proses wrangling yaitu cleaning data, integrasi, eksplorasi, analisis, dan visualisasi. Informasi yang sebelumnya tidak terstruktur dengan baik dapat diolah menjadi data yang layak untuk dianalisis oleh masyarakat dalam membandingkan harga mobil bekas di wilayah DKI Jakarta.

1.2 Rumusan Masalah

1. Bagaimana perbandingan harga mobil bekas di wilayah DKI Jakarta dari berbagai sumber?
2. Apa saja faktor yang mempengaruhi perbedaan harga mobil bekas?
3. Apa keuntungan dari hasil analisis dalam pengambilan keputusan?

1.3 Tujuan

1. Untuk menganalisis dan membandingkan harga mobil bekas dari berbagai sumber informasi di wilayah DKI Jakarta.
2. Untuk mengidentifikasi faktor-faktor yang mempengaruhi perbedaan harga mobil bekas.
3. Untuk menyajikan hasil analisis dalam bentuk visualisasi dan informasi yang mudah dipahami sebagai pertimbangan pengambilan keputusan bagi konsumen dan penjual.

1.4 Manfaat

1. Memberikan informasi perbandingan harga dari beberapa sumber untuk membantu mengetahui harga mobil bekas yang terbaik dan paling relevan di wilayah DKI Jakarta.
2. Menjadi bahan evaluasi dalam menentukan harga yang sesuai dengan kondisi pasar dari beberapa platform online.
3. Membantu konsumen dan penjual memahami faktor-faktor yang mempengaruhi harga mobil bekas.

BAB II
ISI PEMBAHASAN

2.1 Proses Wrangling

2.1.1 Scraping

Pada tahap awal dalam proses scraping, dimulai dari mengambil data dari halaman web. Menggunakan variabel *url* untuk menyimpan alamat web yang berisi daftar mobil bekas di wilayah DKI Jakarta, dilengkapi dengan *page_number* untuk menentukan jumlah data per halaman dan memilih halaman tertentu yang diinginkan. Dengan menggunakan *requests.get()* untuk mengambil isi halaman, lalu diproses oleh *BeautifulSoup* agar mudah dalam proses ekstraksi dan mudah dibaca.

2.1.2 Data Cleaning

Proses data cleaning merupakan bagian untuk merapikan data hasil scraping yang telah dilakukan. Dimulai dari memecah setiap teks pada list mobil menggunakan metode *split()*, dengan adanya pemisahaan ini bertujuan untuk menguraikan data sebelumnya yang masih tergabung dalam satu kalimat seperti brand, tipe, dan tahun, menjadi bagian-bagian yang lebih terstruktur. Setelah data terpecah sendiri-sendiri, tahap berikutnya yaitu membersihkan dan mengekstraksi elemen penting dari *mobil_list*. Pada tahap yang sama, harga juga dibersihkan dengan menghapus format 'Rp' dan merapikan penulisannya. Setelah semua elemen berhasil diekstraksi dan dibersihkan, data tersebut disatukan ke dalam data frame dengan menggunakan *pd.DataFrame* dengan kolom tahun, brand, mobil, dan harga. Proses data cleaning ini menghasilkan data yang lebih rapi, terstruktur, dan siap untuk dianalisis lebih lanjut.

2.1.3 Data Collecting

Proses data collecting dari file CSV yang sudah melewati tahap scraping dan cleaning. File CSV dibaca menggunakan *pd.read_csv()*, dengan memeriksa struktur dan isinya untuk memastikan hasil scraping dan cleaning telah tersimpan dengan baik dan benar. Pada proses ini, dilakukan pengecekan jumlah baris, data setiap kolom, serta memverifikasi bahwa tidak ada nilai yang hilang atau salah dalam proses sebelumnya. Setelah data diproses, dilakukan validasi ulang terhadap variabel penting seperti tahun, brand, model, dan harga untuk memastikan tidak ada data yang

tidak sesuai. Tahap data collecting ini memastikan bahwa data yang sebelumnya diolah kini berada dalam format yang terstruktur, siap dipakai untuk proses analisis lanjutan dan eksplorasi data.

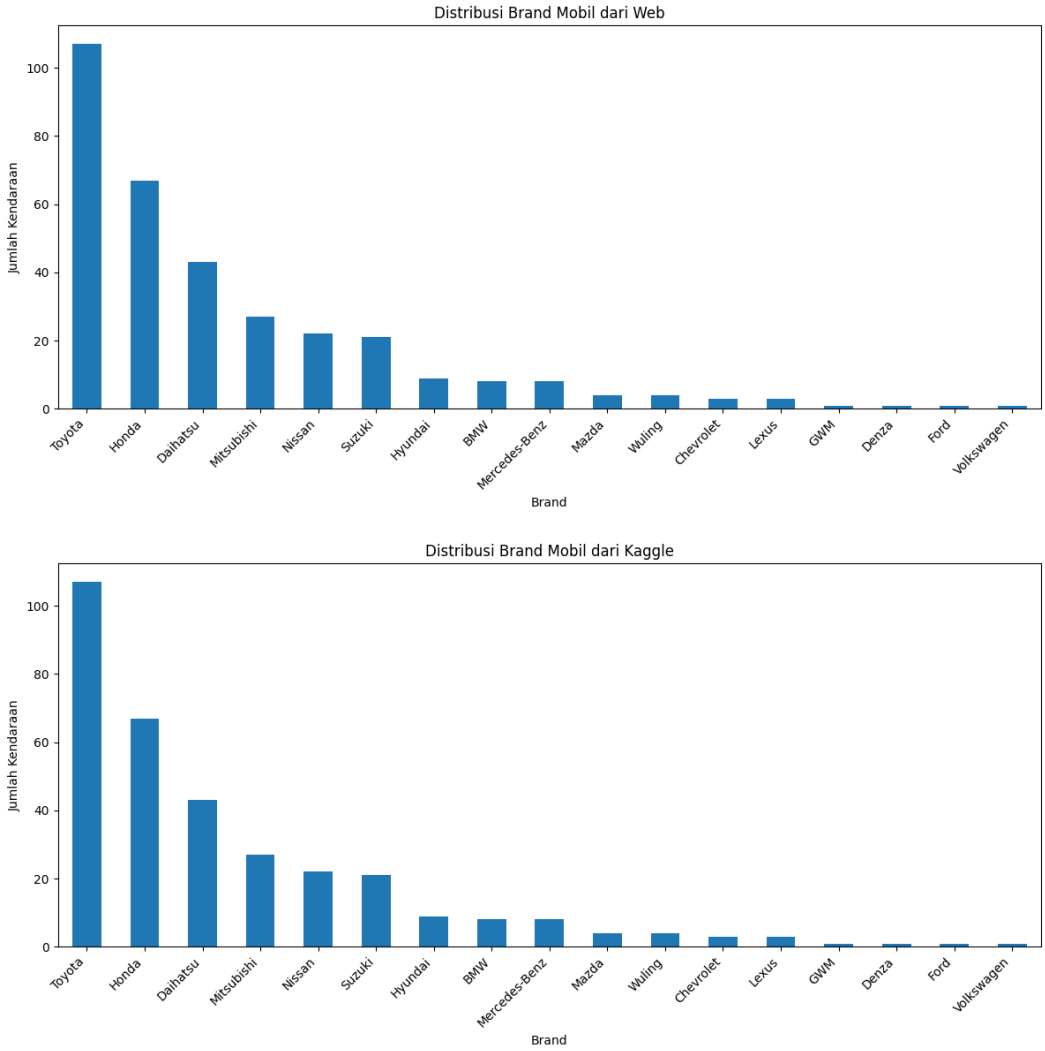
2.1.4 Data Integration

Integrasi data merupakan proses menggabungkan data dari beberapa file untuk menjadi satu dataset yang utuh, terstruktur, dan siap digunakan untuk analisis. Tujuan proses ini yaitu untuk menyatukan data yang tersebar agar tidak terpisah-pisah, menghilangkan duplikasi, menyamakan struktur kolom, dan memastikan setiap data dapat saling melengkapi. Pada proses ini dilakukan melalui perintah *pd.concat([df_carmudi, df_caroline], axis=0)* yang berarti menggabungkan dua data frame. Cara ini dilakukan karena kedua dataset memiliki struktur kolom yang sama, sehingga setiap baris dari berbagai file dapat disatukan menjadi satu tabel besar tanpa mengubah bentuk data. Hasilnya, dijadikan satu dengan *df_web* yang berisi gabungan lengkap dari seluruh baris data mobil yang berasal dari berbagai platform, sehingga dataset menjadi lebih banyak variasi harga mobil.

Selain itu, pada dataset kaggle menggunakan fungi *df_kaggle.drop* dengan menghapus beberapa fitur yang dianggap tidak diperlukan dalam proses analisis data. Fitur yang dihapus seperti foto, tautan, dan pembaruan data biasanya tidak memiliki nilai analisis yang signifikan, terutama ketika fokus analisis adalah pada variabel utama seperti harga, tahun, brand dan tipe mobil. Dengan menghapus beberapa fitur tersebut, data frame menjadi lebih ringkas, lebih mudah diolah, dan meminimalkan gangguan dari informasi yang tidak relevan, sehingga proses pengolahan data dapat berjalan lebih efisien.

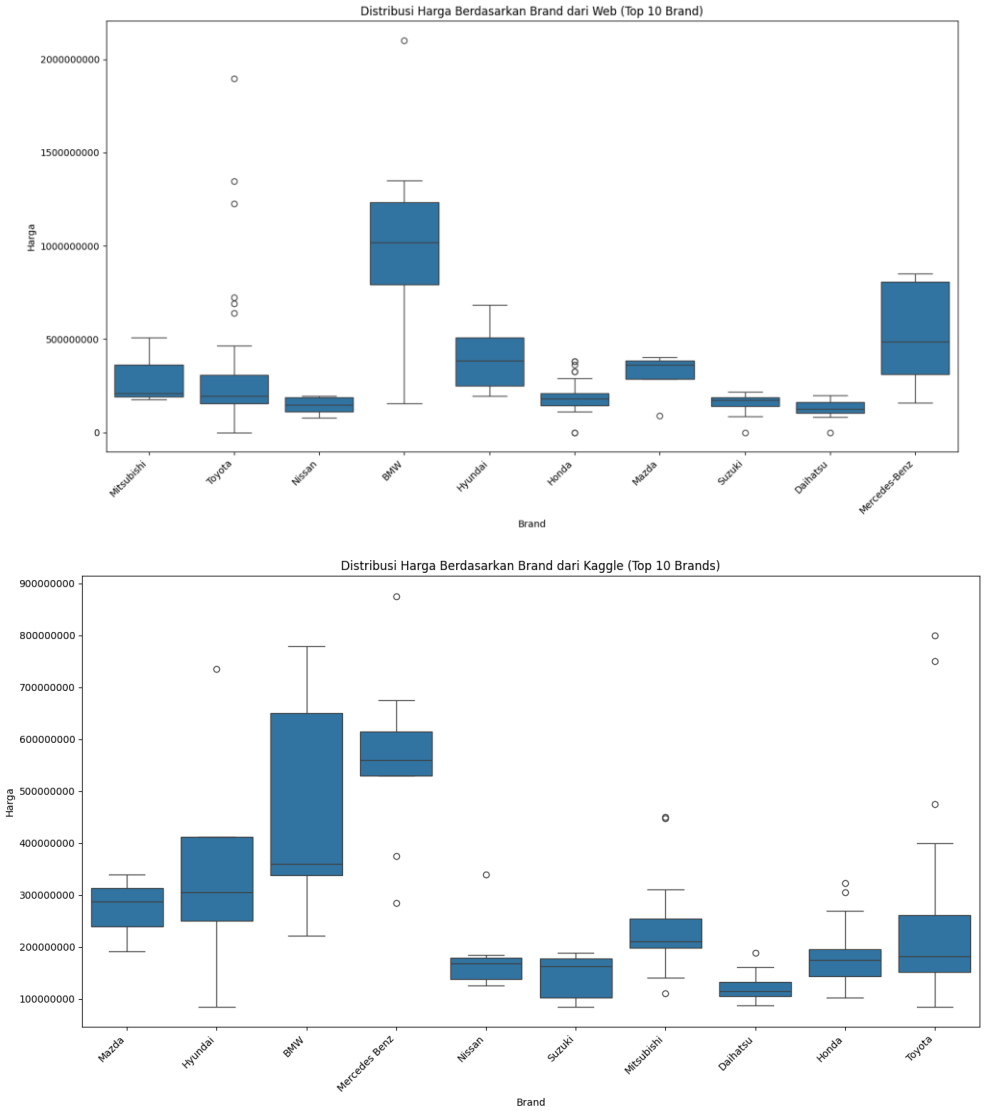
2.1.5 Visualisasi

Visualisasi merupakan proses menyajikan data dalam bentuk grafik, diagram, atau tampilan visual lainnya, yang bertujuan untuk mempermudah pemahaman data dengan menampilkan informasi secara lebih jelas dan lebih mudah dipahami.



Visualisasi Distribusi Brand Mobil Web Scrap & Data Kaggle

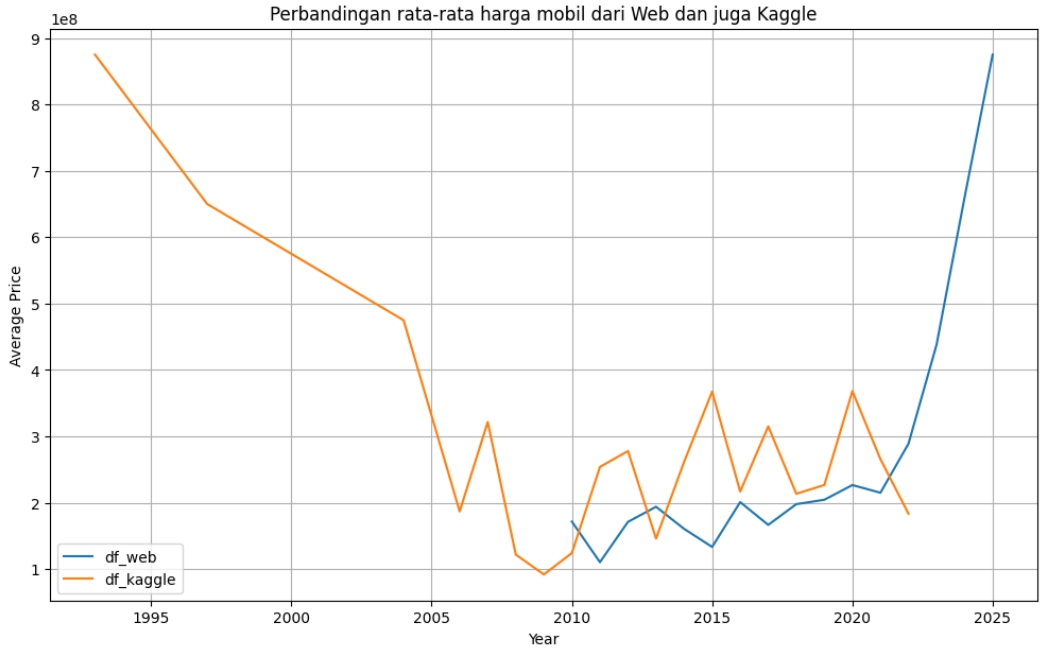
Kedua visualisasi distribusi brand mobil bekas dari dataset hasil scraping dan dataset kaggle menunjukkan pola yang hampir sama. Dimana Toyota menjadi brand dengan jumlah terbanyak dan mendominasi di pasar mobil bekas. Setelah Toyota, brand lain seperti Honda, Daihatsu, Suzuki juga muncul dengan jumlah yang lebih sedikit. Dapat disimpulkan bahwa Toyota menjadi brand yang paling populer dan paling banyak diperdagangkan di ketiga sumber data.



Visualisasi Boxplot Data Web Scrap & Data Kaggle

Pada kedua visualisasi boxplot dari dataset scraping maupun dataset kaggle memiliki tujuan yang sama, yaitu untuk melihat rentang harga setiap brand dan mendeteksi adanya outlier. Pada dataset scraping, terlihat bahwa brand Toyota memiliki outlier harga sekitar 3 miliar, hal tersebut berasal dari model Toyota Land Cruiser sehingga harganya jauh diatas rata-rata mobil Toyota lainnya.

Pada visualisasi boxplot dataset kaggle juga muncul pola outlier, namun terdapat perbedaan yaitu lebih banyak brand yang memiliki rentang harga dan rata-rata yang lebih tinggi dibandingkan data scraping. Hal ini menunjukkan bahwa dataset Kaggle berisi lebih banyak mobil dengan varian harga yang tinggi, sehingga distribusi brand pada harga cenderung lebih besar dan sangat berpengaruh.



Visualisasi Perbandingan Harga Mobil berdasarkan Tahun Keluaran dari Web Scrap & Dataset Kaggle

Berdasarkan perbandingan visualisasi diatas, rata-rata harga pada dataset hasil scraping menunjukkan pola yang relatif stabil hingga tahun-tahun sebelumnya, dimana mobil baru cenderung memiliki harga yang lebih tinggi. Namun, pada tahun 2025 terlihat terjadi lonjakan harga yang sangat ekstrem. Kenaikan ini disebabkan oleh beberapa data mobil premium yang harganya sangat tinggi pada tahun tersebut, sehingga secara otomatis menaikkan rata-ratanya jauh lebih tinggi dibandingkan tahun-tahun sebelumnya. Sementara itu, pada dataset Kaggle terlihat anomali pada tahun 1993 dengan rata-rata harga mendekati 900 juta, yang berasal dari mobil Mercedes-Benz 300GE yang termasuk kendaraan premium dan langka. Dengan demikian, kedua dataset sama-sama menunjukkan adanya nilai ekstrem, tetapi muncul pada tahun yang berbeda dan dipengaruhi oleh brand mobil yang bernilai tinggi pada masing-masing sumber data.

2.1.6 Analisis

Proses analisis diawali dengan memeriksa tipe data pada setiap kolom dan menampilkan beberapa baris awal untuk memastikan bahwa kolom harga telah berformat integer dan data terstruktur dengan benar. Dilanjutkan dengan mengevaluasi hubungan antara tahun mobil dengan harga melalui perhitungan koefisien korelasi serta membuat visualisasi rata-rata harga tiap tahun. Hasil visualisasi dari dataset scraping menunjukkan pola yang awalnya stabil, dimana semakin baru tahun keluaran mobil, maka semakin tinggi rata-rata harga nya. Namun terjadi lonjakan rata-rata harga yang sangat ekstrem pada tahun 2025, karena ada beberapa mobil bekas premium yang bernilai miliaran yang menaikkan rata-rata secara signifikan. Pada dataset kaggle ditemukan lonjakan harga pada tahun tertentu yang disebabkan oleh mobil mewah yaitu Mercedes-Benz 300GE. Hal ini menunjukkan bahwa meskipun tahun keluaran mobil mempengaruhi harga, keberadaan mobil premium pada tahun tertentu dapat menyebabkan lonjakan nilai rata-rata sehingga pola harga tidak selalu konsisten di kedua dataset.

Analisis dilanjutkan dengan mengidentifikasi brand mobil dengan jumlah terbanyak pada masing-masing dataset. Visualisasi distribusi brand menunjukkan bahwa Toyota merupakan brand yang paling dominan pada data scraping maupun juga data kaggle. Setelah itu, visualisasi boxplot dibuat untuk melihat distribusi harga setiap brand serta mendeteksi outlier. Pada dataset web scraping, terlihat bahwa Toyota memiliki outlier dari model Land Cruiser. Pada dataset kaggle juga muncul pola outlier, namun lebih banyak brand yang memiliki rentang harga dan rata-rata lebih tinggi karena adanya banyak mobil premium. Berdasarkan keseluruhan hasil visualisasi, dapat disimpulkan bahwa brand mobil memiliki pengaruh yang lebih besar terhadap variasi harga dibanding dengan tahun keluaran produksi, terutama karena setiap brand mobil memiliki jenis dan kelas yang berbeda-beda dengan harga yang bervariasi.

2.2 Kendala dan Rencana Tindak Lanjut

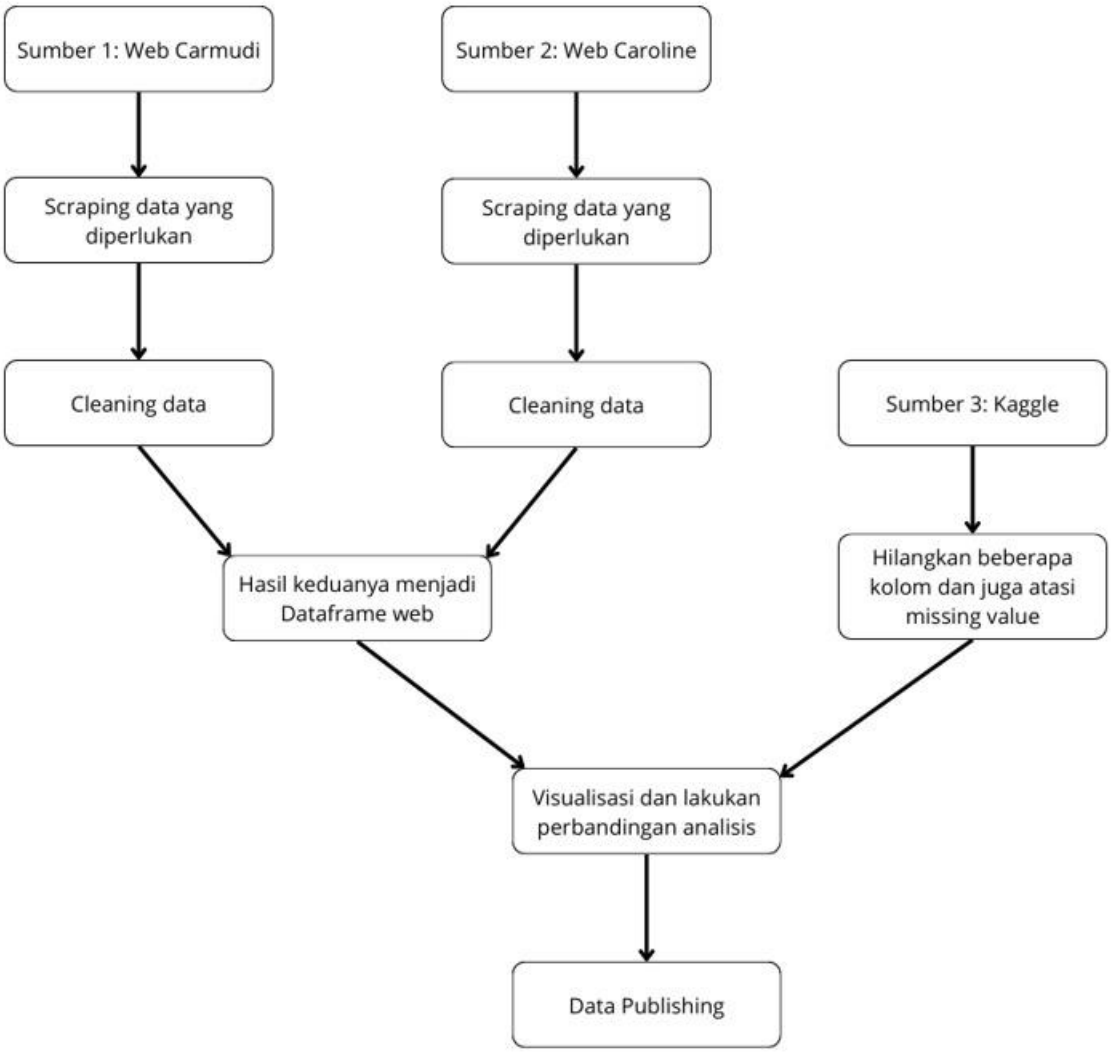
2.2.1 Kendala dari Keseluruhan Proses

Salah satu kendala yang muncul selama keseluruhan proses adalah keterbatasan fitur dan informasi yang disediakan oleh web sumber data. Banyak detail penting yang tidak ditampilkan secara lengkap, seperti spesifikasi kendaraan, kondisi mobil, riwayat pemakaian, dan lain sebagainya, sehingga informasi yang dapat dikumpulkan dari proses scraping menjadi kurang mendalam. Keterbatasan ini membuat analisis tidak bisa dilakukan secara lebih detail, karena beberapa fitur yang sebenarnya berpengaruh terhadap harga mobil tidak tersedia dalam data yang diambil. Akibatnya, hasil analisis menjadi terbatas pada informasi dasar yang disediakan oleh situs tersebut.

2.2.2 Rencana Analisis Lanjutan Setelah Proses Wrangling

Setelah proses wrangling, analisis lanjutan dapat difokuskan pada beberapa aspek penting. Seperti melakukan analisis perbandingan antar-platform untuk mengidentifikasi perbedaan harga antara Carmudi, Caroline, dan dataset Kaggle pada mobil yang memiliki spesifikasi serupa, sehingga dapat mengetahui perbedaan harga atau kecenderungan tertentu pada masing-masing platform. Selain itu, mungkin dapat dilakukan analisis tren tahun ke tahun untuk melihat bagaimana perkembangan harga mobil bekas dari seiring berjalannya waktu dan perkembangan zaman, apakah terdapat pola kenaikan atau penurunan yang konsisten.

2.3 Dokumentasi Pipeline



BAB III
PENUTUP

3.1 Kesimpulan

Berdasarkan seluruh rangkaian proses yang dimulai dari scraping, cleaning, collecting, integrasi data, visualisasi, hingga analisis, dapat disimpulkan bahwa data harga mobil bekas di wilayah DKI Jakarta memiliki keragaman yang cukup tinggi antar platform. Proses wrangling berhasil mengubah data mentah yang tidak terstruktur menjadi dataset yang rapi dan layak untuk dianalisis. Analisis menunjukkan bahwa faktor brand memiliki pengaruh besar terhadap harga, dan mobil dengan brand yang premium memiliki rentang harga yang lebih tinggi dan lebih bervariasi.

Dengan adanya beberapa visualisasi, seperti distribusi brand yang menunjukkan bahwa Toyota merupakan brand yang paling dominan dan paling banyak muncul, visualisasi tersebut menggambarkan popularitasnya di pasaran mobil bekas. Visualisasi harga berdasarkan tahun keluaran menunjukkan pola kenaikan harga yang relatif stabil, dimana mobil dengan tahun yang lebih baru memiliki harga lebih tinggi. Dari visualisasi boxplot dapat dilihat bahwa variasi harga per brand jauh lebih signifikan dibandingkan variasi harga berdasarkan tahun produksi. Beberapa brand memiliki rentang harga yang tinggi serta outlier yang mencolok.

Dapat disimpulkan bahwa brand mobil sangat berpengaruh besar terhadap harga mobil dibandingkan dengan tahun keluaran, karena setiap brand memiliki kelas dan kategori mobil yang berbeda-beda. Secara keseluruhan, analisis ini menghasilkan informasi yang relevan dan dapat dimanfaatkan sebagai dasar pertimbangan dalam pengambilan keputusan, baik bagi konsumen maupun penjual yang ingin memahami kondisi harga pasar mobil bekas secara lebih objektif.

3.2 Referensi

[1] Attaqi, M. I. D., & Wibowo, J. S. (2025). *Prediksi Harga Mobil Bekas Berdasarkan Tipe Penjual dan Jenis Kendaraan Menggunakan Regresi Linier*. Department of Informatics, Universitas Stikubank (UNISBANK) Semarang, Indonesia.

[2] Hasibuan, E., & Karim, A. (2022). *Implementasi Machine Learning untuk Prediksi Harga Mobil Bekas dengan Algoritma Regresi Linear Berbasis Web*. Jurnal Ilmiah KOMPUTASI, 21(4).

[3] Syukur, M. A. A., & Faisal, M. (2023). *Penerapan Model Regresi Linear untuk Estimasi Mobil Bekas Menggunakan Bahasa Python*. EULER: Jurnal Ilmiah Matematika, Sains dan Teknologi, 11(2).

3.3 Lampiran

Link Google Collab: Project_DataWrangling_Kelompok12.ipynb

Link GitHub: https://github.com/fawwazazri/ProjekUAS-DatWrang

3.4 Kontribusi

Fawwaz Azri Manshur Prasetya [24031554074]	● Melakukan web scraping untuk mengumpulkan data dari berbagai platform. ● Melakukan pembersihan data (cleaning) agar data siap digunakan. ● Melakukan penggabungan dataset (data integration) dari beberapa sumber. ● Menyusun laporan presentasi ● Membuat visualisasi data untuk menampilkan pola dan tren. ● Melakukan analisis statistik atau eksploratif terhadap dataset. ● Menyusun pipeline proses data yang terstruktur.
Meisya Natasafira [24031334174]	● Menyusun laporan yang sistematis dari tahap awal hingga akhir. ● Menyusun laporan presentasi yang sistematis dari tahap awal hingga akhir. ● Menganalisis seluruh tahapan proses wrangling. ● Melakukan pembersihan data (cleaning) agar data siap digunakan. ● Mengidentifikasi faktor yang mempengaruhi harga mobil bekas, dengan menjelaskan metode, kendala, dan keputusan yang diambil selama proses. ● Menyusun pola harga berdasarkan tahun produksi atau brand sekaligus kesimpulan berdasarkan hasil analisis.