Exploratory Data Analysis on the Auto Mobile Dataset

Report

Ntokozo Mazibuko

Contents

Introduction

This EDA explores the Automobile dataset, which contains technical specs, efficiency metrics, and descriptive features for a variety of passenger cars. The goal is to uncover how specifications influence price and fuel economy, compare patterns across manufacturers and body styles, and identify key trade-offs, extremes, and outliers.

The analysis involves:

- Data preparation: cleaning, formatting, and deriving a combined mpg metric.
- Exploration: statistical summaries, correlations, and targeted visualizations.
- Segmentation: comparisons by manufacturer and body style, with deeper looks at hatchbacks and high-displacement models.

The findings aim to provide evidence-based insights for pricing benchmarks, product positioning, and consumer guidance, while setting the stage for deeper modeling or decision support.

Data Cleansing

To turn the raw file into a reliable analysis table, we applied a set of defensible, auditable steps. The goal was to remove formatting noise, enforce correct data types, and guarantee that every row used in charts/stats is complete.

Standardize and normalize inputs

- Missing markers: Converted the placeholder "?" to proper missing values (NaN) across all columns so pandas could detect and handle them uniformly.
- Text hygiene: Trimmed leading/trailing spaces in all text fields to avoid spurious categories (e.g., "sedan " vs "sedan"). Case and spelling in manufacturer names were preserved as supplied.

Enforce correct data types

- Numeric casting: Safely coerced numeric-like columns—price, engine-size, horsepower, city-mpg, highway-mpg, curb-weight, and others—to numeric types. Any non-parsable entries became NaN (later removed), preventing silent string math.
- Categorical typing: Converted descriptive fields (make, body-style, fuel-type, drive-wheels, engine-type, num-of-cylinders, fuel-system, etc.) to categorical dtype. This speeds up grouping and yields cleaner summaries.
- Note: num-of-cylinders remains a labeled category (“four”, “six”, ...) rather than an integer. This preserves semantics for reporting; it can be mapped to numbers later for modeling if needed.
- Derived metric: Added combined_mpg = (city-mpg + highway-mpg) / 2 to compare fuel efficiency with a single, intuitive measure.

De-duplication and completeness

- Duplicate rows: Removed 0 exact duplicate rows (no effect on sample size).
- Complete-case filter: Dropped 45 rows with any remaining missing values (≈ 22% of the raw 205 records). This avoids imputation assumptions in a small dataset and ensures all downstream stats/plots are computed on fully observed cases.

Sanity checks (post-clean)

- No missing values remain in the analysis table.
- Ranges are plausible for key fields: prices and engine specs are positive; mpg values are within automotive norms.
- Types are consistent: numeric columns are truly numeric; descriptor columns are categorical.

Rationale and trade-offs

- We favoured complete-case analysis over imputation to keep the EDA transparent and assumption-light. The trade-off is a smaller sample (from 205 to 160 rows), but the benefit is cleaner, more trustworthy comparisons—especially for price and mpg.
- We did not minorize or cap outliers at this stage; the EDA intentionally shows the full spread, including luxury/performance extremes.
- Provenance and reproducibility
- The curated dataset was saved as automobile_clean.csv, and the cleaning logic is captured in the notebook, so results are reproducible end-to-end.

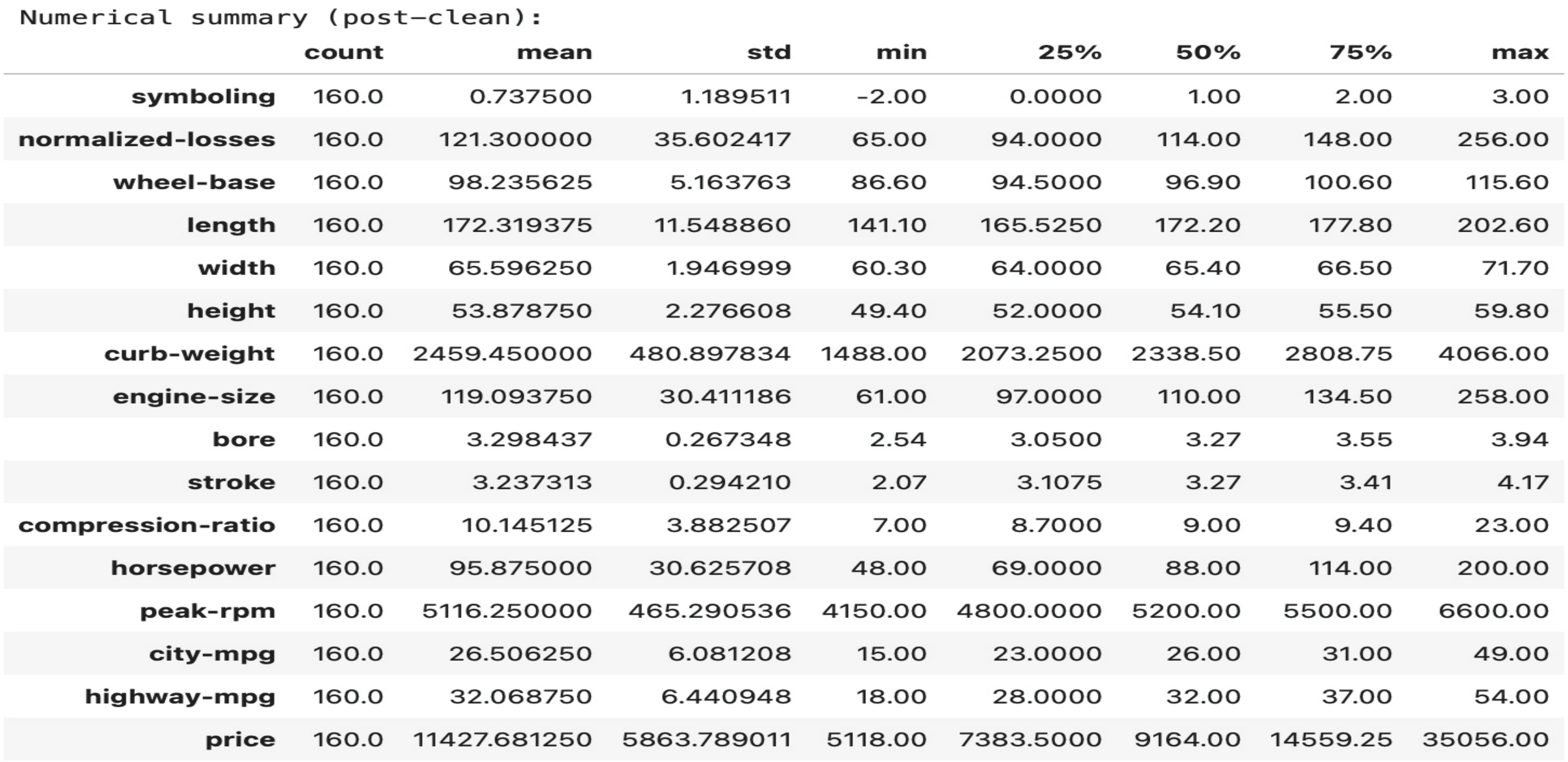
Final analysis table: 160 rows × 27 columns

- (26 original fields + 1 derived: combined_mpg; 0 duplicates; 45 rows removed for missing data; 0 missing values remain.)

Missing Data

- Missing markers: Converted the placeholder "?" to proper missing values (NaN) across all columns so pandas could detect and handle them uniformly.
- 45 rows removed for missing data

Summary Statistic after data Cleansing

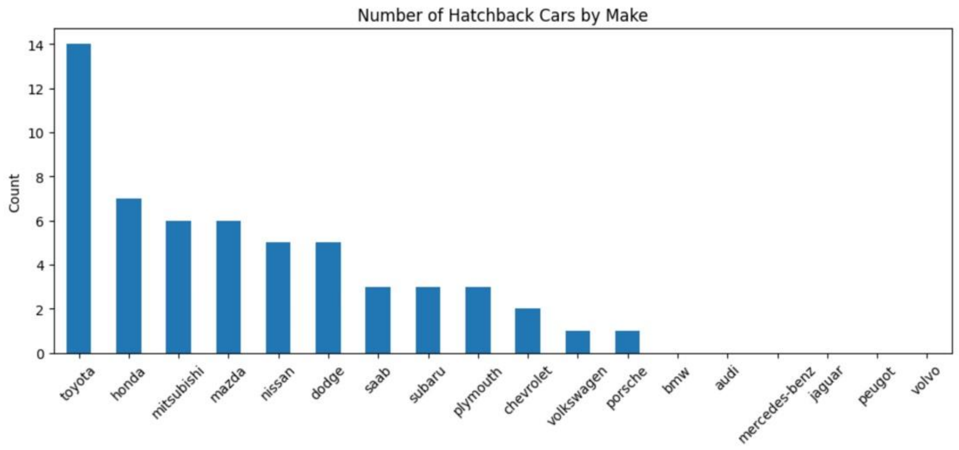


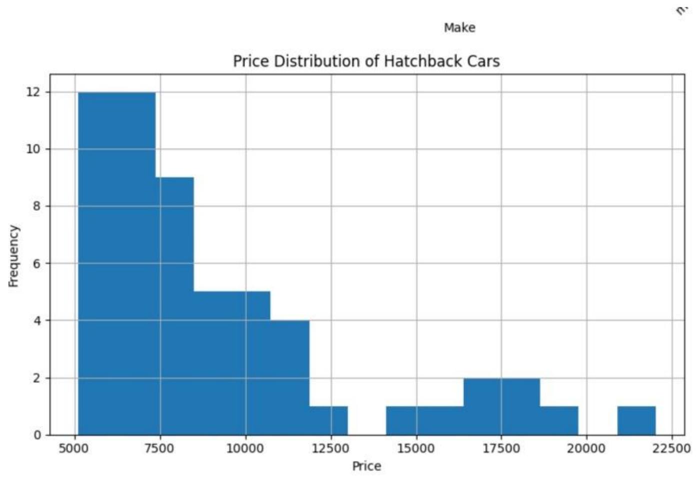
```
Numerical summary (post-clean):
```

	count	mean	std	min	25%	50%	75%	max
symboling	160.0	0.737500	1.189511	-2.00	0.0000	1.00	2.00	3.00
normalized-losses	160.0	121.300000	35.602417	65.00	94.0000	114.00	148.00	256.00
wheel-base	160.0	98.235625	5.163763	86.60	94.5000	96.90	100.60	115.60
length	160.0	172.319375	11.548860	141.10	165.5250	172.20	177.80	202.60
width	160.0	65.596250	1.946999	60.30	64.0000	65.40	66.50	71.70
height	160.0	53.878750	2.276608	49.40	52.0000	54.10	55.50	59.80
curb-weight	160.0	2459.450000	480.897834	1488.00	2073.2500	2338.50	2808.75	4066.00
engine-size	160.0	119.093750	30.411186	61.00	97.0000	110.00	134.50	258.00
bore	160.0	3.298437	0.267348	2.54	3.0500	3.27	3.55	3.94
stroke	160.0	3.237313	0.294210	2.07	3.1075	3.27	3.41	4.17
compression-ratio	160.0	10.145125	3.882507	7.00	8.7000	9.00	9.40	23.00
horsepower	160.0	95.875000	30.625708	48.00	69.0000	88.00	114.00	200.00
peak-rpm	160.0	5116.250000	465.290536	4150.00	4800.0000	5200.00	5500.00	6600.00
city-mpg	160.0	26.506250	6.081208	15.00	23.0000	26.00	31.00	49.00
highway-mpg	160.0	32.068750	6.440948	18.00	28.0000	32.00	37.00	54.00
price	160.0	11427.681250	5863.789011	5118.00	7383.5000	9164.00	14559.25	35056.00

Hatchback Analysis

Hatchbacks in this dataset occupy a value-driven niche in the market. They cluster around lower price points, often pairing compact engines with above-average fuel efficiency, making them particularly attractive for budget-conscious buyers and city commuters.





Composition & Brand Presence

- Dominated by mass-market brands such as Toyota, Nissan, Honda, and Volkswagen, ensuring wide availability in entry-level and mid-tier trims.
- Luxury and premium brands are largely absent, contributing to the segment's generally modest pricing profile.

Pricing Profile

- Price distribution is right-skewed, concentrated in the budget to lower mid-range bracket.
- Occasional sport-oriented or turbocharged models emerge as high-price outliers, but these remain the exception rather than the norm.

Chart Insights

- Hatchbacks by Make (bar chart): A small group of manufacturers dominate production, highlighting brand leaders in the segment.
- Hatchback Price Distribution (histogram): Majority fall in the affordable range, with outliers clearly identifiable.

Key Takeaways

- For buyers prioritizing affordability and efficiency, hatchbacks deliver the strongest value proposition.
- Those seeking premium features or high performance may need to look beyond this segment—or accept reduced fuel efficiency within it.
- Market concentration among leading brands suggests better availability, choice, and price negotiation potential.

Next steps

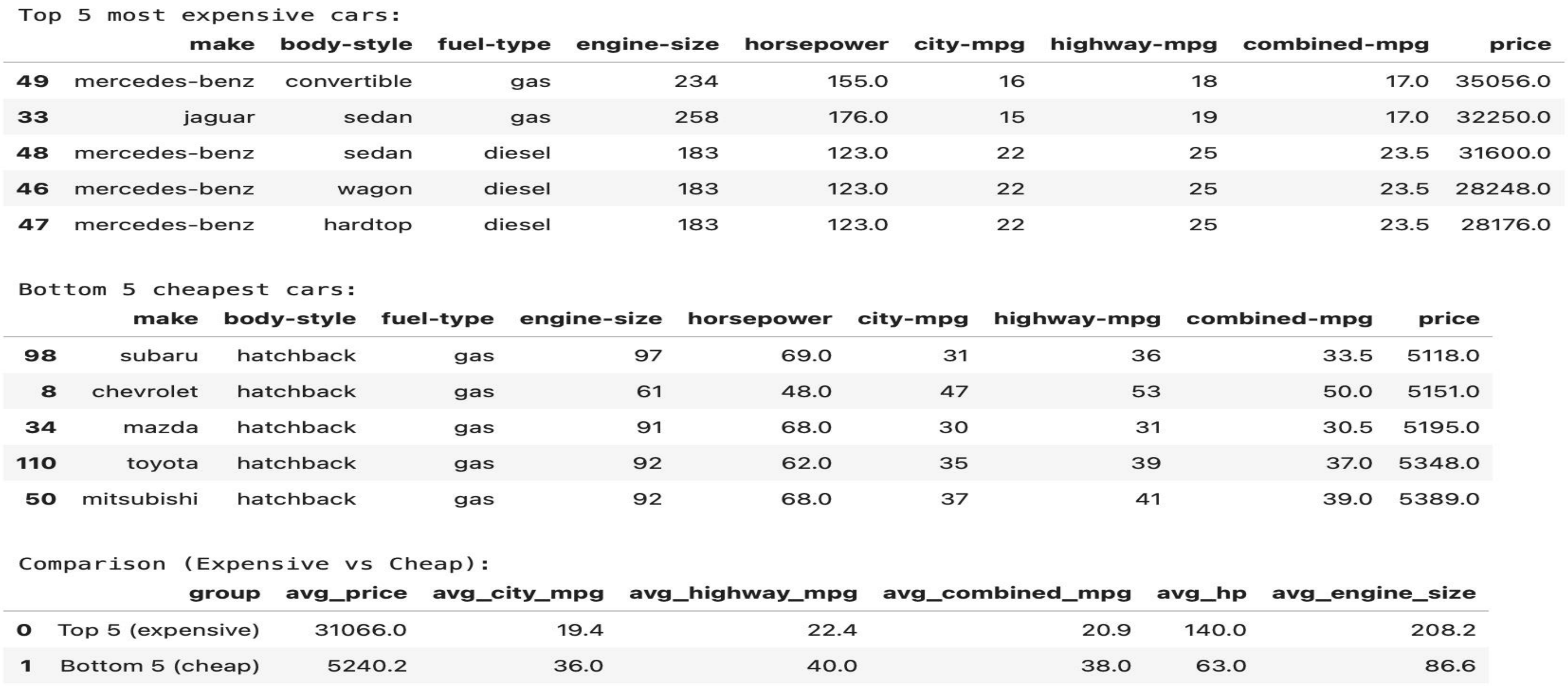
- Build a price prediction model (e.g., log-price with regularized regression or tree-based methods) and validate with cross-validation.
- Test robustness: compare results with imputed versions of the data; consider log transforms for skewed variables.
- Add richer features (e.g., power-to-weight, cylinder count numeric mapping) and quantify marginal effects on price and MPG.

Overall, the analysis consistently shows: more performance → higher price, lower MPG; compact designs and economy-focused brands → better efficiency and value.

Data Stories and Visualisation

Top 5 most expensive and bottom priced 5 cars comparison Analysis

This comparison spotlights the price extremes in the dataset to make trade-offs tangible. We selected five highest priced and five lowest priced cars (post data cleansing) and compare their key specs. The below tables highlights top and bottom prices cars.



Top 5 most expensive cars:

	make	body-style	fuel-type	engine-size	horsepower	city-mpg	highway-mpg	combined-mpg	price
49	mercedes-benz	convertible	gas	234	155.0	16	18	17.0	35056.0
33	jaguar	sedan	gas	258	176.0	15	19	17.0	32250.0
48	mercedes-benz	sedan	diesel	183	123.0	22	25	23.5	31600.0
46	mercedes-benz	wagon	diesel	183	123.0	22	25	23.5	28248.0
47	mercedes-benz	hardtop	diesel	183	123.0	22	25	23.5	28176.0

Bottom 5 cheapest cars:

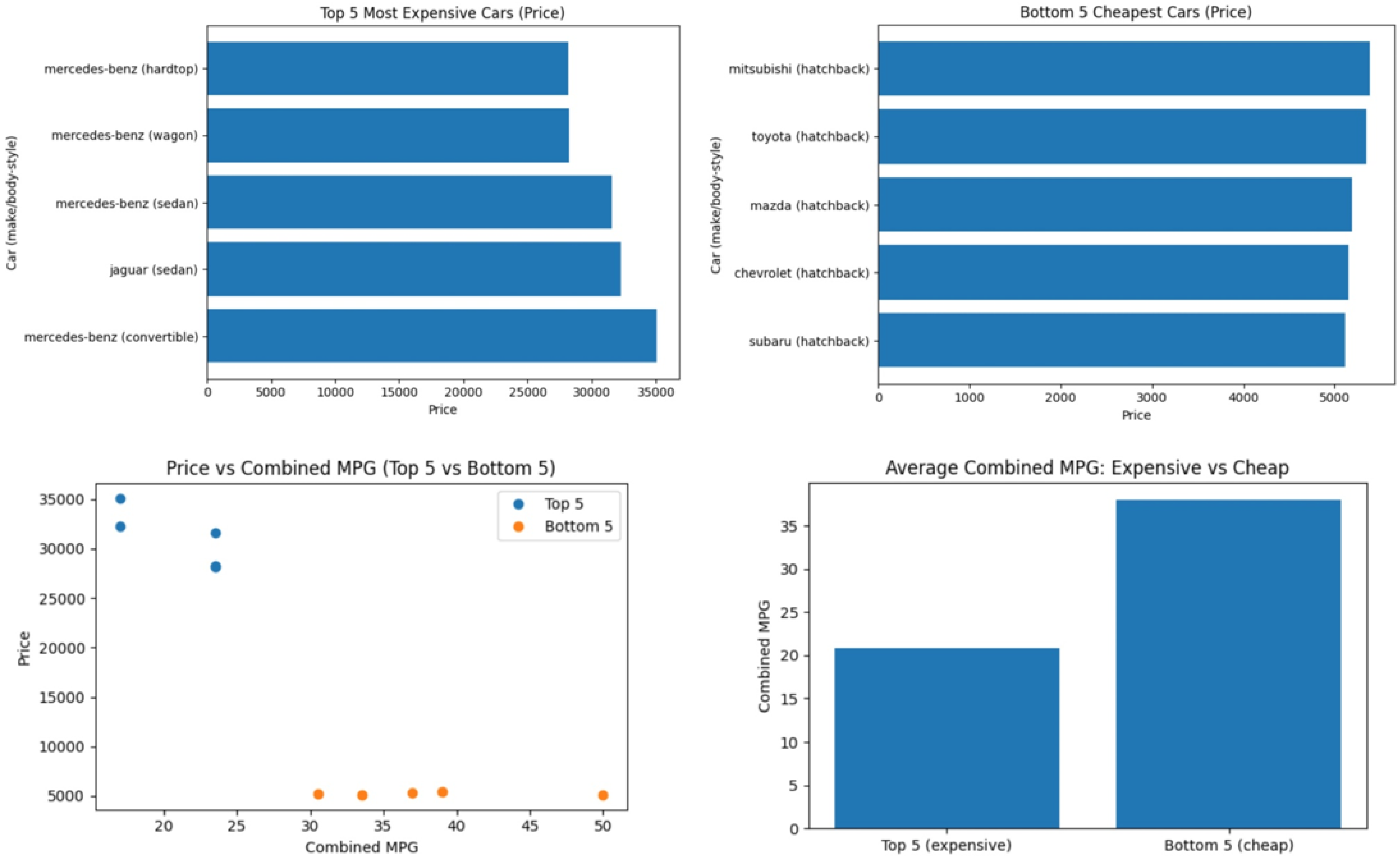
	make	body-style	fuel-type	engine-size	horsepower	city-mpg	highway-mpg	combined-mpg	price
98	subaru	hatchback	gas	97	69.0	31	36	33.5	5118.0
8	chevrolet	hatchback	gas	61	48.0	47	53	50.0	5151.0
34	mazda	hatchback	gas	91	68.0	30	31	30.5	5195.0
110	toyota	hatchback	gas	92	62.0	35	39	37.0	5348.0
50	mitsubishi	hatchback	gas	92	68.0	37	41	39.0	5389.0

Comparison (Expensive vs Cheap):

	group	avg_price	avg_city_mpg	avg_highway_mpg	avg_combined_mpg	avg_hp	avg_engine_size
0	Top 5 (expensive)	31066.0	19.4	22.4	20.9	140.0	208.2
1	Bottom 5 (cheap)	5240.2	36.0	40.0	38.0	63.0	86.6

Key observations:

- It quickly reveals how performance and luxury features (bigger engines, higher horsepower) typical drives prices up.
- It also shows the efficiency divided at the lower end: cheaper cars tend to deliver better fuel economy.


Top 5 Most Expensive Cars (Price)
Car (make/body-style)
mercedes-benz (hardtop)
mercedes-benz (wagon)
mercedes-benz (sedan)
jaguar (sedan)
mercedes-benz (convertible)
0
5000
10000
15000
20000
25000
30000
35000
Price
Bottom 5 Cheapest Cars (Price)
Car (make/body-style)
mitsubishi (hatchback)
toyota (hatchback)
mazda (hatchback)
chevrolet (hatchback)
subaru (hatchback)
0
1000
2000
3000
4000
5000
Price
Price vs Combined MPG (Top 5 vs Bottom 5)
Top 5
Bottom 5
Price
35000
30000
25000
20000
15000
10000
5000
20
25
30
35
40
45
50
Combined MPG
Average Combined MPG: Expensive vs Cheap
Combined MPG
35
30
25
20
15
10
5
0
Top 5 (expensive)
Bottom 5 (cheap)

Key observations:

Top 5 Most Expensive Cars

- Dominated by Mercedes-Benz (multiple body styles) and one Jaguar sedan.
- Prices range from roughly $28k to $35k, reflecting a strong luxury segment presence.

Bottom 5 Cheapest Cars

- Exclusively hatchbacks from brands like Mitsubishi, Toyota, Mazda, Chevrolet, and Subaru.
- Prices cluster tightly around $5k–$5.3k, showing little spread within the budget segment.

Price vs. Combined MPG (Top 5 vs Bottom 5)

- Clear inverse relationship:
  - Expensive models average low combined MPG (~18–25), indicating a performance or luxury bias over efficiency.
  - Cheapest models achieve high MPG (~30–50), underscoring their economy focus.

Average Combined MPG: Expensive vs Cheap

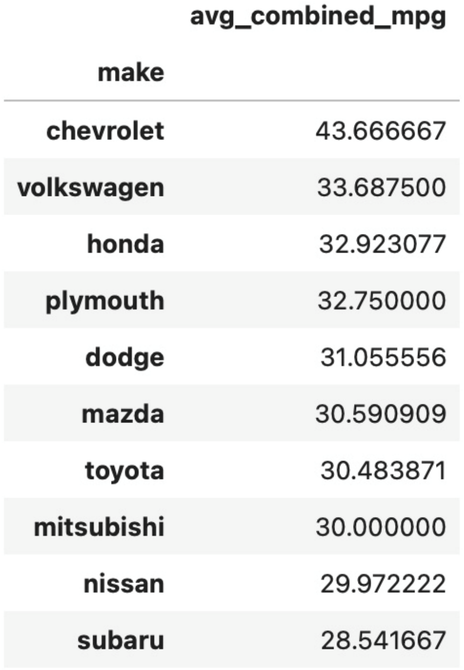
- Budget cars average nearly double the MPG of the luxury segment (≈ 38 vs 20).
- This stark gap reinforces the trade-off between performance/features and fuel efficiency.

Overall Takeaway:

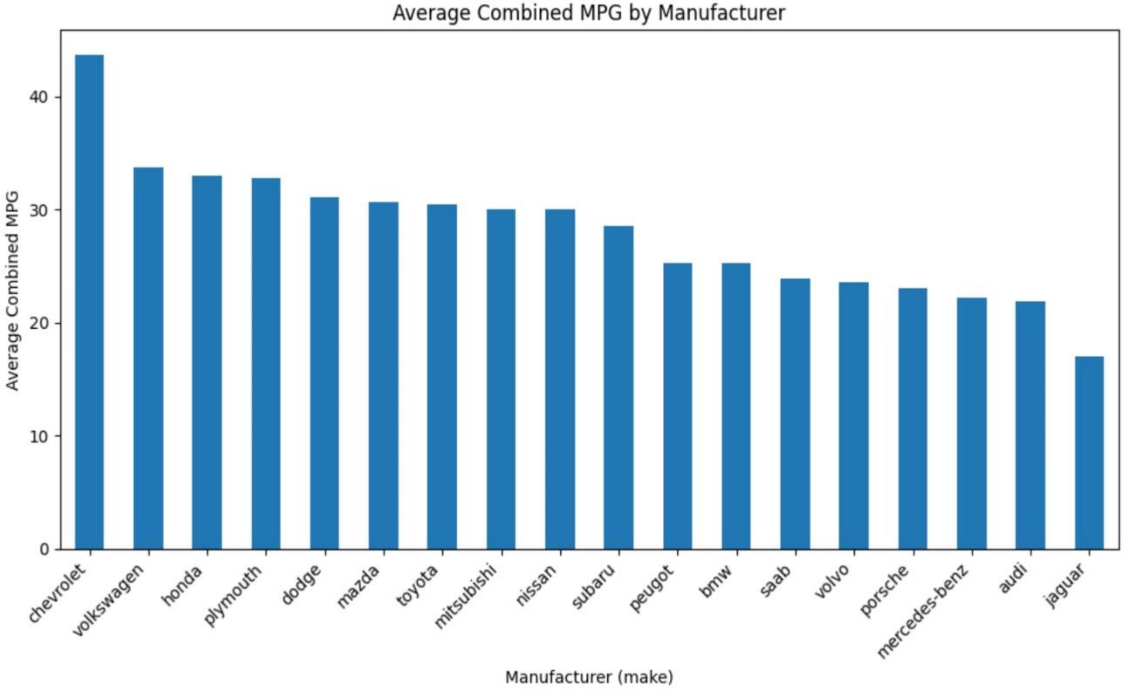
The data underscores a strong divide between the luxury and budget segments. High-priced models, dominated by Mercedes-Benz and Jaguar, sacrifice fuel efficiency for performance and premium features, while low-cost hatchbacks deliver exceptional MPG. The nearly twofold efficiency gap between the two groups highlights the inherent trade-off between affordability and efficiency versus luxury and performance.

Mile per Gallon Analysis

Fuel economy is a core lens for the comparing vehicles, so we analyse city, highway and combined MPG to understand efficiency patterns across makes and body style. We profile the distribution of MPG, identify bands with the highest average efficiency and examine how MPG trades off against engine size, horsepower and price.



	avg_combined_mpg
make	
chevrolet	43.666667
volkswagen	33.687500
honda	32.923077
plymouth	32.750000
dodge	31.055556
mazda	30.590909
toyota	30.483871
mitsubishi	30.000000
nissan	29.972222
subaru	28.541667



Key observations:

Top Performers in Efficiency

- Chevrolet stands out with the highest average combined MPG (43.67), significantly ahead of all other brands.
- Volkswagen (33.69) and Honda (32.92) also deliver strong fuel efficiency, both well above the overall dataset average.

Consistent Economy Segment

- Brands like Plymouth, Dodge, Mazda, Toyota, and Mitsubishi cluster in the 30–33 MPG range, indicating reliable economy-focused offerings.

Lower Efficiency Luxury/Performance Segment

- Premium and performance-oriented brands (BMW, Porsche, Mercedes-Benz, Audi, Jaguar) fall below 27 MPG, with Jaguar posting the lowest at 20.0 MPG.
- This reflects the expected trade-off between performance features and fuel efficiency.

Clear Market Segmentation

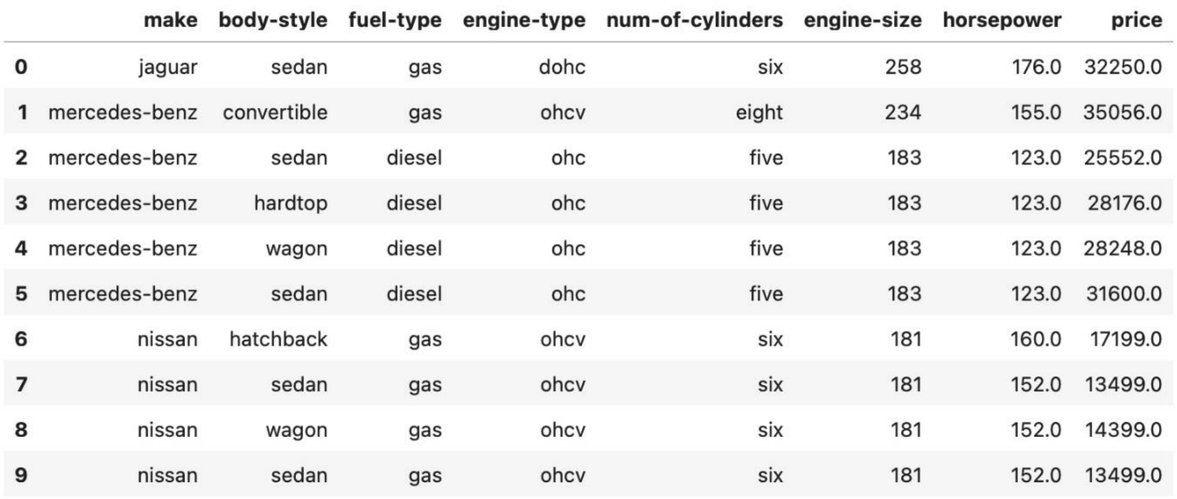
- The bar chart visually emphasizes a steep drop in MPG after Chevrolet, highlighting the unique outlier status of its efficiency.
- The gap between economy-focused and performance-focused brands is pronounced, suggesting distinct market strategies.

Overall takeaway:

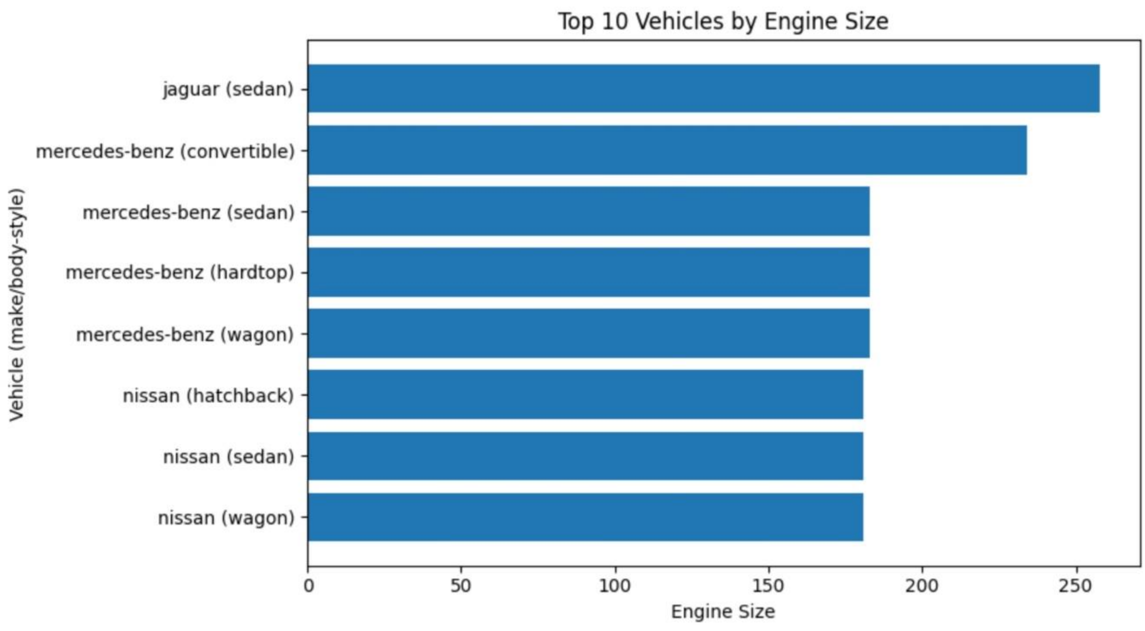
Fuel efficiency results reveal clear market segmentation, with Chevrolet emerging as a standout outlier far ahead of all competitors. Economy-focused brands maintain consistent MPG performance in the low 30s, while luxury and performance brands trade efficiency for power, clustering below 27 MPG. The sharp visual drop after Chevrolet underscores its exceptional position, while the divide between economy and performance segments reflects deliberate brand strategy differences.

Large Engine Capacity Analysis

This section highlights the performance end of the dataset by ranking vehicles by engine-size (displacement), our proxy for engine capacity. Cars at the top of this list typically pair higher horsepower and premium pricing with lower fuel economy, reflecting a clear performance–efficiency trade-off. We present a Top 10 by engine size chart to show which makes and body styles dominate this extreme and to anchor later comparisons on price and MPG.



	make	body-style	fuel-type	engine-type	num-of-cylinders	engine-size	horsepower	price
0	jaguar	sedan	gas	dohc	six	258	176.0	32250.0
1	mercedes-benz	convertible	gas	ohcv	eight	234	155.0	35056.0
2	mercedes-benz	sedan	diesel	ohc	five	183	123.0	25552.0
3	mercedes-benz	hardtop	diesel	ohc	five	183	123.0	28176.0
4	mercedes-benz	wagon	diesel	ohc	five	183	123.0	28248.0
5	mercedes-benz	sedan	diesel	ohc	five	183	123.0	31600.0
6	nissan	hatchback	gas	ohcv	six	181	160.0	17199.0
7	nissan	sedan	gas	ohcv	six	181	152.0	13499.0
8	nissan	wagon	gas	ohcv	six	181	152.0	14399.0
9	nissan	sedan	gas	ohcv	six	181	152.0	13499.0



Key observations:

Largest Engine Size:

- The Jaguar sedan has the largest engine size at 258, followed by the Mercedes-Benz convertible at 234.

Dominance of Mercedes-Benz:

- Mercedes-Benz appears five times in the top 10 list, with multiple body styles (convertible, sedan, hardtop, wagon), mainly using diesel engines with an engine size of 183.

Nissan's representation:

- Nissan has four entries (hatchback, sedan, and wagon), all with engine sizes of 181, showing consistency in their higher-capacity engines.

Fuel type trends:

- Gas engines dominate, but several top Mercedes-Benz models use diesel fuel.

Price correlation:

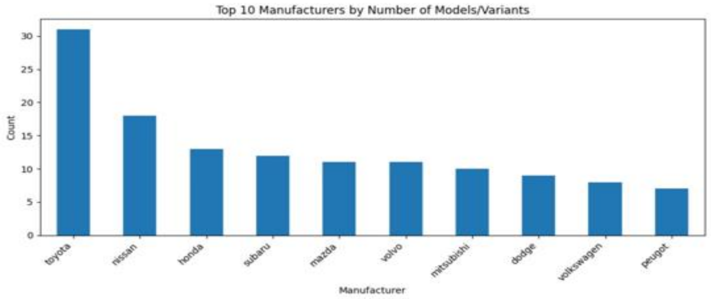
- Higher engine sizes generally align with higher prices, though Nissan models are significantly cheaper than Jaguar and Mercedes-Benz, indicating brand and market positioning differences.

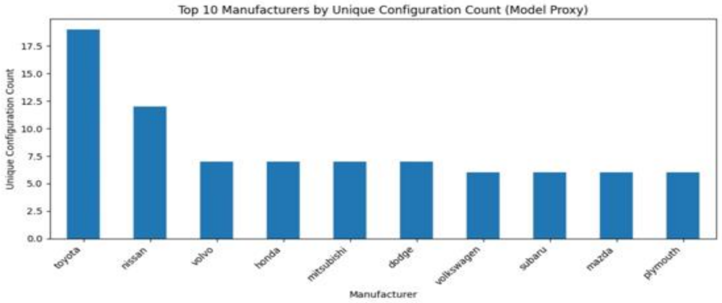
Overall takeaway:

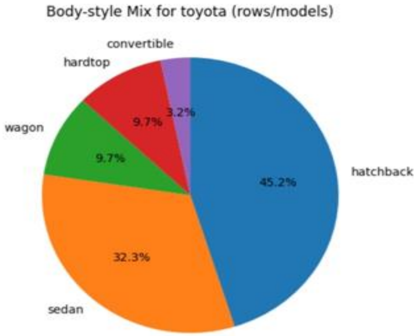
The data highlights a clear performance hierarchy, with Jaguar holding the largest engine size and Mercedes-Benz dominating the high-capacity segment through diverse body styles and a notable use of diesel engines. Nissan shows consistent engine sizing at a more affordable price point, underscoring differences in brand positioning. Overall, larger engine sizes tend to correspond with higher prices, and the visual gap in the chart reinforces the outlier status of Jaguar and the Mercedes-Benz convertible compared to the rest of the field.

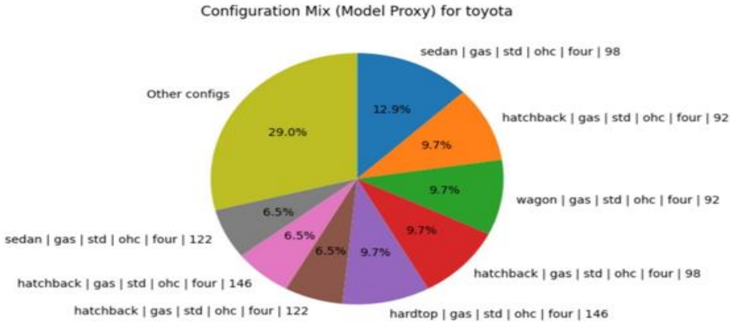
Manufacturer with most car models Analysis

To answer this, we count the number of entries per manufacturer (make) in the cleaned dataset, treating each row as a model/variant. This highlights which brands have the broadest representation and helps contextualize brand-level averages elsewhere in the EDA. (As a robustness check, we also report a unique-configuration proxy—counting distinct combinations like make + body-style + fuel-type + cylinders + engine-size—to approximate distinct models.)









Key observations:

Top 10 manufacturers by total models/variants (row count)

- This represents how many entries each brand has in the dataset.
- Toyota leads decisively with about 31 entries, followed by Nissan with around 18, then Honda, Subaru, Mazda, Volvo, Mitsubishi, Dodge, Volkswagen, and Peugeot.
- Larger counts for Toyota and Nissan mean their averages are more statistically stable. Note this reflects *dataset coverage*, not market share.

Top 10 manufacturers by unique configurations (model proxy)

- Number of distinct spec combinations per brand, such as body style, fuel type, aspiration, engine type, cylinders, and engine size.
- Toyota remains first with approximately 19 unique configurations, followed by Nissan with about 12, while most other brands cluster around 6–7.
- The difference between total rows and unique configurations suggests Toyota and Nissan have many repeated trims or variants rather than completely different powertrains.

Toyota body-style distribution (pie chart)

- Hatchbacks make up 45.2%, sedans 32.3%, wagons 9.7%, hardtops 9.7%, and convertibles 3.2%.
- Toyota's dataset presence leans toward compact and value-oriented body types, which helps explain higher MPG and lower average prices compared to more performance-focused brands.

Toyota configuration distribution (model-proxy pie chart)

- Several small four-cylinder gasoline configurations each account for about 6.5–12.9%, with "Other configurations" comprising a significant 29% tail.
- There is broad variety within Toyota's small gas engine specs and trims, with many minor variations beyond the top combinations.

Overall takeaway

Toyota dominates the dataset in both volume and variety, especially within hatchbacks and sedans. To compare brands fairly, rely on per-brand metrics like average price and MPG, and keep in mind that dataset coverage does not equal market share.

Conclusion

This EDA turned a raw automobile spec/price file into a clean, analysis-ready dataset (160 rows × 27 columns, with a derived combined_mpg metric and 0 missing values) and used it to explain how specs map to cost and efficiency.

Key Learnings

- Price is driven by performance. It rises clearly with engine size, horsepower, and (often) curb-weight; fuel economy moves in the opposite direction. The correlation matrix and scatter plots show a strong performance–efficiency trade-off.
- Body style matters. Convertibles/coupes sit at the high end of average price; sedans/wagons land mid-range; hatchbacks cluster as the best value/efficiency segment.
- Manufacturers differ in profile. Economy-focused brands top the average combined MPG ranking, while performance/luxury makers fall lower due to larger, more powerful engines. Coverage is uneven—Toyota has the deepest representation—so brand averages should be read with sample size in mind.
- Extremes illustrate the trade-off. The Top 5 most expensive cars pair big engines and high horsepower with lower MPG; the Bottom 5 cheapest sacrifice power but deliver notably better efficiency.
- Largest engines = performance end. Vehicles with the biggest displacement anchor the high-price, low-MPG corner of the space—useful reference points when benchmarking.
- Segment snapshot—Hatchbacks. Dominated by mass-market brands, hatchbacks are affordable and fuel-efficient; a few turbo/sport trims break the pattern but are rare.