Honda Car Data Scraping and Analysis Report

Source Data Provider: AckoDrive

Data Scraped: Honda Car Models and Specifications

Date of Scraping: November 25, 2025

1. Executive Summary

This report documents the process and results of scraping current and discontinued Honda car model data from the AckoDrive website. The objective was to extract key specifications—Model Name, Fuel Type, Transmission, Price Range, and status—and organize this information into a structured dataset. The scraping process yielded **7 unique model entries** (which included duplicates of the most recent models and separate entries for older discontinued versions) and demonstrated the challenge of cleanly extracting data from unstructured text blocks. The final cleaned dataset is ready for further analysis, such as trend tracking or competitive benchmarking.

2. Methodology: Web Scraping Process

The data was collected using a Python script, leveraging core libraries for web interaction and data processing.

2.1 Tool Stack

- **Request Handling:** requests library to fetch the HTML content of the target URL (https://ackodrive.com/collection/honda+cars/).
- **Parsing and Extraction: BeautifulSoup** for navigating the HTML structure and identifying car model elements.
- **Data Structure: pandas** for handling, cleaning, and exporting the final structured data.

- **Text Processing: Python's re *(regular expressions)*** for pattern matching to extract structured details (like fuel type and transmission) from unstructured text blocks associated with each car model card.

2.2 Extraction Challenges

The primary challenge encountered was the lack of clear, consistent CSS selectors for individual data fields (e.g., fuel type, transmission). Instead of using specific HTML tags, the script relied on **regular expressions** to search for keywords within the large text block surrounding each model name. This approach, while effective in this case, highlights the need for continuous maintenance if the source website's text layout changes.

3. Data Structure and Cleaning

3.1 Initial Raw Data Fields

The scraping script successfully extracted the following raw fields for each model card:

Field Name	**Description**
Brand	Always "Honda".
Model	The specific model name (e.g., "Honda City," "Honda Elevate").
Variants	Placeholder, mostly captured as "N/A" due to extraction method limitations.
Fuel	Extracted fuel types (e.g., "Hybrid • Petrol Manual • Automatic").

Transmission	Extracted transmission types (e.g., "Manual • Automatic").
price_range_raw	Raw text price range, captured as "N/A" in this sample run.
is_discontinued	Boolean flag, derived from searching for the keyword "Discontinued".
detail_url	Direct link to the model's detail page on AckoDrive.

3.2 Data Cleaning and Enrichment

The following cleaning steps were applied to prepare the data for analysis:

1. **Date Stamping:** An operational column, scrape_date, was added, set to 2025-11-25.
2. **Price Normalization:** A function was implemented to parse the price_range_raw string (e.g., "₹13.9 lakh – ₹23.9 lakh") into two dedicated numeric fields: price_min_lakh and price_max_lakh. *Note: In the sampled output, this field was consistently 'N/A', resulting in None values for min/max prices.*
3. **Deduplication:** Although not explicitly shown as a step in the provided code, cleaning would require deduplicating entries like the two identical "Honda City" entries found, or deciding whether to retain entries for older models (e.g., "Honda Amaze (2021-2024)").

4. Analysis of Scraped Data

The final cleaned dataset contains **7 rows** of unique Honda car model entries (excluding the duplicates found in the raw scrape).

4.1 Car Model Summary

The following table summarizes the key specifications identified:

Index	Model Name	Fuel Options	Transmission Options	Status (Discontinued)	Detail URL
0	Honda City	Hybrid • Petrol	Manual • Automatic	No	.../honda-city/
1	Honda Elevate	Petrol	Manual • Automatic	No	.../honda-elevate/
2	Honda Amaze	Petrol	Manual • Automatic	No	.../honda-amaze/
3	Honda Amaze (2021-2024)	Petrol	Manual • Automatic	No	.../honda-amaze-2021-2024/
4	Honda WR-V (2020-2023)	Petrol • Diesel	Manual	No	.../honda-wr-v-2020-2023/

5	Honda Jazz (2020-2023)	Petrol	Manual • Automatic	No	.../honda-jazz-2020-2023/
6	Honda City (2020-2023)	Hybrid • Petrol • Diesel	Manual • Automatic	No	.../honda-city-2020-2023/

4.2 Key Findings

- **Current Flagship Models:** The primary current models appear to be the **Honda City, Honda Elevate, and Honda Amaze**.
- **Fuel Diversity:**
  - The Honda City is offered with the most diverse fuel options, including **Hybrid** and **Petrol**.
  - The older Honda City (2020-2023) also included a **Diesel** option, suggesting a shift away from diesel in newer models.
- **Transmission Standard:** Almost all models offer both **Manual** and **Automatic** transmission options, demonstrating Honda's commitment to catering to different driving preferences.
- **Discontinued Models:** No models were explicitly tagged as "Discontinued" (is_discontinued = False) in this run, despite several entries clearly being past models (e.g., 2020-2023 variants). This suggests the explicit "Discontinued" tag was either missing on the source page or not correctly captured by the heuristic extraction logic.

5. Conclusion and Recommendations

The scraping notebook successfully retrieved structured data from a semi-structured web source, demonstrating a robust use of regex-based heuristics when standard element selectors are unavailable. The resulting dataset provides a clear overview of Honda's model lineup, fuel, and transmission offerings.

Recommendation for Future Work:

1. **Refine Price Extraction:** Investigate why the price_range_raw field was captured as "***N/A***" and adjust the parent block traversal or regex to reliably capture the price information, which is critical for market analysis.
2. **Deduplication and Filtering:** Implement a formal deduplication step to retain only the most current version of each model (e.g., keep "Honda City" but filter out "Honda City (2020-2023)") for a cleaner final dataset.
3. **Variant Analysis:** Update the scraping logic to successfully capture the number of variants, moving the variants column away from its current "***N/A***" status.